Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Cancer Informatics

INsPeCT: INtegrative Platform for Cancer Transcriptomics

Piyush B. Madhamshettiwar[1,2], Stefan R. Maetschke[1,2], Melissa J. Davis[1,2], Antonio Reverter[3] and Mark A. Ragan[1,2]

[1]The University of Queensland, Institute for Molecular Bioscience, St. Lucia, Brisbane, Queensland, Australia. [2]Australian Research Council Centre of Excellence in Bioinformatics, St. Lucia, Brisbane, Queensland, Australia. [3]CSIRO Animal, Food and Health Sciences, St. Lucia, Brisbane, Queensland, Australia.

ABSTRACT: The emergence of transcriptomics, fuelled by high-throughput sequencing technologies, has changed the nature of cancer research and resulted in a massive accumulation of data. Computational analysis, integration, and data visualization are now major bottlenecks in cancer biology and translational research. Although many tools have been brought to bear on these problems, their use remains unnecessarily restricted to computational biologists, as many tools require scripting skills, data infrastructure, and powerful computational facilities. New user-friendly, integrative, and automated analytical approaches are required to make computational methods more generally useful to the research community. Here we present INsPeCT (INtegrative Platform for Cancer Transcriptomics), which allows users with basic computer skills to perform comprehensive in-silico analyses of microarray, ChIP-seq, and RNA-seq data. INsPeCT supports the selection of interesting genes for advanced functional analysis. Included in its automated workflows are (i) a novel analytical framework, RMaNI (regulatory module network inference), which supports the inference of cancer subtype-specific transcriptional module networks and the analysis of modules; and (ii) WGCNA (weighted gene co-expression network analysis), which infers modules of highly correlated genes across microarray samples, associated with sample traits, eg survival time. INsPeCT is available free of cost from Bioinformatics Resource Australia-EMBL and can be accessed at http://inspect.braembl.org.au.

KEYWORDS: cancer, systems biology, transcriptomics, transcriptional module networks, microarray, ChIP-seq, RNA-seq

CITATION: Madhamshettiwar et al. INsPeCT: INtegrative Platform for Cancer Transcriptomics. *Cancer Informatics* 2014:13 59–66 doi: 10.4137/CIN.S13630.

RECEIVED: November 1, 2013. **RESUBMITTED:** January 8, 2014. **ACCEPTED FOR PUBLICATION:** January 8, 2014.

ACADEMIC EDITOR: J.T. Efird, Editor in Chief

TYPE: Original Research

FUNDING: Access to TRANSFAC was provided by Queensland Facility for Advanced Bioinformatics through Australian Research Council grant LE098933. PBM, SRM, MJD, and MAR acknowledge the support of Australian Research Council grants CE0348221, DP110103384, and LE0989334.

COMPETING INTERESTS: Author(s) disclose no potential conflicts of interest.

CORRESPONDENCE: m.ragan@uq.edu.au

Background

Over the last decade, the biological sciences have been revolutionized by the development of high-throughput technologies for the study of gene expression, initially microarrays, and then next-generation sequencing (NGS). One result is the enormous quantity of data now publicly available for different cancer types, from independent researchers and consortia such as The Cancer Genome Atlas (TCGA)[1] and the International Cancer Genome Consortium (ICGC),[2] via online data repositories including Sequence Read Archive (SRA),[3] gene expression omnibus (GEO),[4] and ArrayExpress.[5,6] With the continuing improvement in technologies and decreasing costs, transcriptome analysis is becoming routine in cancer research. Analysis and interpretation of large data, however, remain an ongoing challenge.[7]

Many tools are available for transcriptome analysis,[8–11] but their use too often remains limited to those skilled in bioinformatics because these tools have been developed as stand-alone packages, written in different programming languages with only command-line control, intended for a very specific use and/or without support for sharing input and output data with other programs.[12–15] For instance, Bowtie[13] and Samtools[15] are powerful tools for processing raw sequencing reads, but are difficult for biologists to install, configure, and use, and

require memory-efficient high-performance computational resources.

User-friendly web interfaces have been developed,[9,10,16,17] particularly for initial data processing, file-format conversion, and downstream functional analysis. However, most such interfaces provide only a limited set of tools without automated procedures for data import; they impose file-format and data-size restrictions, are difficult to set up, are expensive to access, and/or have been developed commercially. For example, the MEME Suite[18] provides powerful tools for motif-based sequence analysis and is available free of cost to the academic community, but analyzing ChIP-seq reads and preparing data for input to MEME can be challenging for a biologist with limited informatics skills, and it can likewise be difficult to format and redirect their output to programs outside the suite, eg for druggability analysis. To address these barriers to adoption, we have developed INsPeCT. INsPeCT consists of two central types of components: primary modules for high-throughput data analysis, and secondary modules for gene-list analysis and automated network inference.

Three primary analysis modules are provided for microarray, ChIP-seq, and RNA-seq data. In the microarray data-analysis framework, for example, a researcher can import data available online, or upload data from his or her own computer; carry out differential expression analysis and use the list of differentially expressed genes to infer a gene regulatory network (GRN); conduct pathway, druggability, and survival analysis; and/or redirect interesting genes to secondary functional analyses.

Two secondary analysis modules are provided: one covers gene-list analysis, the other provides automated workflows for module-based GRN inference and analysis. The gene-list analysis framework can be used with any gene list of interest; for a given list of genes, it supports analyses including promoter-sequence extraction, druggability analysis, functional enrichment analysis, and transcription factor binding-site over-representation analysis. In the automated workflow framework, researchers can take advantage of our novel analytical framework regulatory module network inference (RMaNI) for automated identification of cancer subtype-specific transcriptional module networks, and execute the widely used weighted gene co-expression network analysis (WGCNA) method, to identify the module network associated with a clinical variable of interest, for example overall survival, relapse-free survival, or metastasis-free survival.

Individual components in INsPeCT have previously been benchmarked and/or compared with earlier tools; below, we provide leading literature citations. For example, the RMaNI framework within INsPeCT supports the inference of modules and their condition-specific regulators, based on the publicly available learning module networks (LeMoNe) algorithm, which was benchmarked against the state-of-the-art genomica and other methods.[19,20] Similarly, differential gene-expression analysis methods edgeR and DEseq, used in the RNA-seq framework, were compared against each other.[21] All components in INsPeCT can be executed quickly in a fully automated way, without need for specialized informatics skills. For the options selected, INsPeCT automatically prepares the input data for downstream analyses. It provides result tables and figures in user-friendly formats, and stores these outputs as individual R objects so the user can reproduce that part of the analysis or perform additional procedures locally without repeating the complete set of operations. Overall, INsPeCT is a computational system biology tool to integrate the analysis of multiple high-throughput data types with advanced downstream functional analysis and network inference for cancer transcriptomics. Our focus on usability and accessibility will make these advanced tools available to a new audience of research scientists.

Implementation

Figure 1 shows the system architecture diagram of INsPeCT. INsPeCT is a web- based application with Drupal CMS front end and server side tools developed mainly by integrating R[22] and Bioconductor[23] packages, Python, Matlab, API-based scripts, web-automation, and web-scraping functions. The web interface to individual programs has been created using Rwui,[24] a Java-based application that uses the Apache Struts framework. The complete application is running on a high-performance computing cluster. The platform integrates over 110 publicly available R, Bioconductor, and custom packages and functions for data import, processing, analysis, integration, and visualization. All packages are currently running under R version 2.15.2 and can be easily updated with newer versions of R. INsPeCT supports up to 2 GB of data upload in any module. INsPeCT is available free of cost from http://inspect.braembl.org.au. User manual and test datasets can

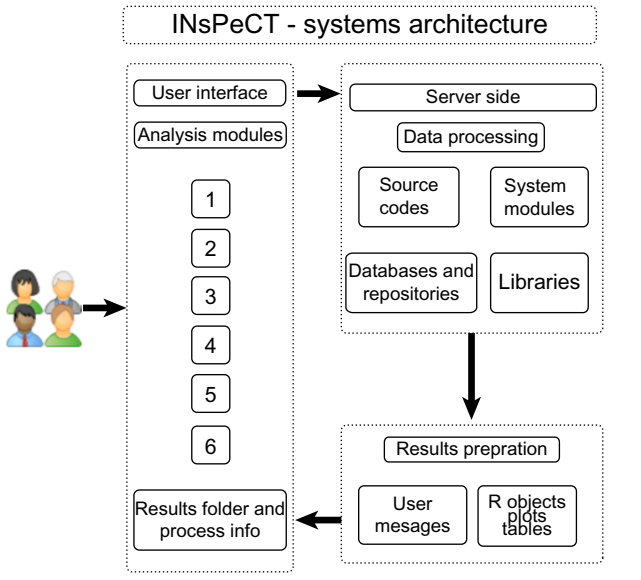


Figure 1. Systems architecture diagram of INsPeCT.

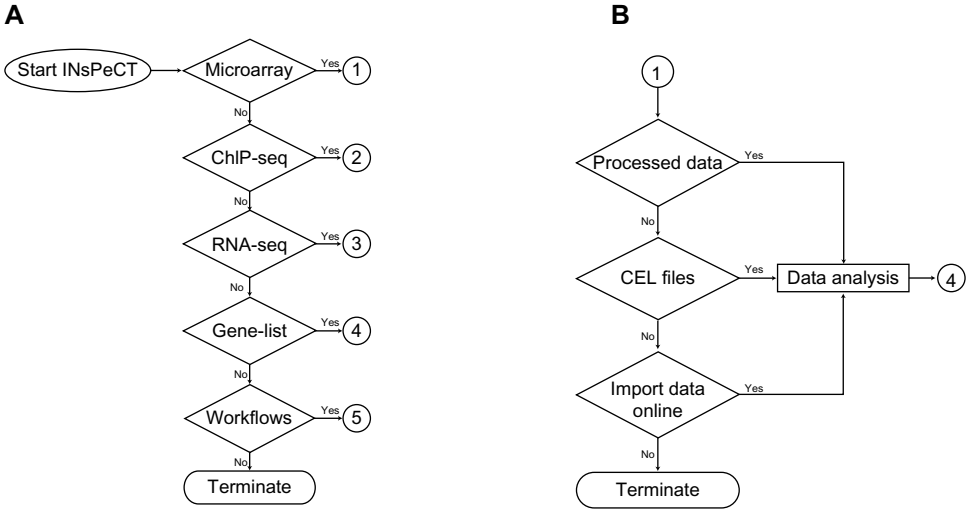
be accessed from the INsPeCT homepage. For the sample datasets provided, approximate run times for complete modular workflows are: microarray and WGCNA framework for less than two hours, ChIP-seq and RNAseq datasets for six hours, Gene-lists framework for less than one hour, and RMaNI for four hours, depending on the machine load.

INsPeCT: Structure and Functionalities

INsPeCT is organized into five main modules for analysis of microarray (Fig. 2B), ChIP-seq (Fig. 2C), RNA-seq raw (Fig. 2C), and processed data (Fig. 2D), and secondary analysis of gene lists and automated workflows (Fig. 2E). The main functionalities in each section are described in detail below.

A. Microarray data. The microarray analysis framework of INsPeCT (Fig. 2B) provides a complete analysis workflow, from raw data import and differential analysis to gene network inference and pathway analysis. Below we describe the critical steps and functions. For the different options selected, INsPeCT automatically prepares the input data for downstream analyses.

Data upload/import and processing. INsPeCT allows the uploading of raw or processed data arising from different types



E

4

Analysis functions				
Microarray	ChIP-seq	RNA-seq	Genelist	Automated workflows
- Data import and pre-processing - Differential expression analysis - Probe summarization - Data selection - Trend analysis - Survival analysis - Regulatory impact factor analysis - Signalling pathway impact analysis - Differential pathway analysis - modified cancer outlier profile analysis (mCOPA) - Gene regulatory network inference - Clustering - Redirect to genelist analysis	- File type conversions- SRA to FastQ to SAM to BAM - Detailed qualicty control - Reference mapping and alignment - Peak calling and peak annotaion - TSS plots - Enrichment analysis of peaks - Extract peak and promoter sequences - Analysis of peaks using MEME suite motif discovery and comparison motif enrichment analysis - Differential binding analysis - Redirect to genelist analysis	- Normalisation of read counts - Annotation - Filtering - Dispersion estimation - Differential analysis using DEseq or edgeR or both - Redirect to genelist analysis	- Druggability analysis - Gene set enriohment analysis (GSEA) - Gene set enriohment disequilibrium analysis (GSEDA) - David enrichment analysis - Exract promoter sequences for genes - Detect CpG islands - Sequences analysis using MEME suite Motif discovery Motif comparison Motif enrichment analysis Motif functional anlysis - TFBS overrepresentation analysis	- Regulatory module networference and analysis (MRaNI) - Weighted gene Co-expression network inference and anlaysis (WGCNA)

Figure 2. Schematic representation of INsPeCT. (**A**) Overall organization of INsPeCT for data processing; (**B**) microarray data analysis workflow; (**C**) ChIP-seq and RNA-seq raw data analysis workflow; (**D**) RNA-seq processed data analysis workflow; and (**E**) functions available to analyze data in each module.

of Affymetrix chips and automated online data import from the GEO or ArrayExpress databases. The user enters the dataset accession ID to import it using the online data importer function. Uploaded or imported raw data will be processed for background correction and normalization using the R package limma. We provide five widely used normalization methods: RMA, GCRMA, MAS5, PLIER, and dChip. Multiple probes can be summarized to one gene using either the coefficient of variation or the maximum average expression methods.[25]

Differential expression analysis. In INsPeCT, differentially expressed genes can be detected using two widely used methods, LIMMA[26] and SAM.[27] We declare a differential gene expression significant if the Benjamini—Hochberg (BH) adjusted *P*-value is at most 0.05. The result of differential expression analysis is provided as a comma separated value (CSV) file, and the data passed to downstream analysis modules.

Trend analysis. Users can identify genes with a consistent increase or decrease in median expression (monotonic trend) across multiple conditions. This analysis is performed using a trend analysis function based on the Jonckheere—Terpstra (JT) test, and using the SAGx package.[28] If this option selected, INsPeCT will automatically prepare the input data for trend analysis.

Regulatory impact factor analysis. To identify the transcription factors potentially regulating differentially expressed genes across two conditions, regulatory impact factor analysis (RIF)[29] can be used. We developed an R function to implement the RIF FORTRAN code. This analysis can be useful to priorities transcription factors or microRNAs based on their regulatory potential with respect to a given set of genes, or to infer a regulatory network. If this option selected, INsPeCT will automatically prepare the input data for RIF.

GRN inference. INsPeCT supports several tools for the inference GRNs including mutual information-based methods (relevance networks,[30] ARACNE,[31] CLR,[32] and MRNET),[33] gene correlation (Pearson, Spearman, and Kendall-tau), partial correlation and information theory (PCIT),[34] and regression trees (GENIE3).[35] These tools originate from the MINET, PCIT, and GENIE3 R packages.[35–37] The Transfac2009.4[38] and Tcof-DB[39] databases are used for transcription factor information. For a comprehensive comparative evaluation of these GRN inference methods, please see Ref. 40.

Clustering. Microarray samples can be clustered according to their expression similarity using the CLUES, KMEANS, PAM, AGNES, FANNY, SOTA, and MCLUST methods, available from the CLUES, MCLUST, and cluster R packages.[41–43] The user can compare the performance of these tools on their data to select the most-appropriate method.

Signaling pathway impact analysis. A signaling pathway impact analysis algorithm is available in the SPIA package.[44] It uses differential gene expression and log fold changes together with signaling pathway topology from the KEGG database release 64.0[45] to identify the pathways that are perturbed in an experiment.

Differential gene set analysis. Differential gene set analysis can be performed using one of the broad gene-set collections from MsigDB.[46] INsPeCT currently supports all six gene set types (positional, curated, motifs, computational, GO, and onogenic signature gene sets). INsPeCT uses the sigpathway R package[47] for differential gene set analysis.

Modified cancer outlier profile analysis (mCOPA). mCOPA[48] is a new tool for the exploration of cancer datasets and discovery of new cancer subtypes, and can be combined with pathway and functional analysis approaches to discover mechanisms underpinning heterogeneity in cancers. The biology explored by outlier analysis can differ from that uncovered in differential expression or variance analysis.[48]

Redirecting analysis to genelist analysis framework. Lists of interesting genes identified in microarray analysis (eg, as differentially expressed) can be redirected to the genelist analysis framework (Fig. 2D) for additional functional analysis. Genelist analysis framework is described in more detail below (section E).

B. ChIP-seq data. The ChIP-seq analysis framework of INsPeCT (Fig. 2C) provides a complete analysis pipeline from raw data import to motif discovery and analysis. Below we describe the critical steps and functions. For the different options selected, INsPeCT automatically prepares the input data for downstream analyses.

File-format conversions, mapping reads to a reference genome, and quality control. Mapping reads to a reference genome is the first step in analysis of ChIP-seq data; we implement Bowtie for this purpose. The standard file-format for raw ChIP-seq reads used for input to Bowtie is FASTQ, which we then convert to SAM, BAM, and sorted BAM formats for further analyses. For data that are available in SRA format, we provide functionality to convert from SRA to FASTQ. We also provide functionality to upload mapped reads in any of the SAM, BAM, or sorted BAM file-formats. We implement sratoolkit[49] and samtools[15] for these file-format conversions. After mapping reads, we process the data for quality control using the FastQC tool.[50] INsPeCT provides an interactive HTML-based report to review the results of comprehensive quality control checks.

Peak calling. Peak calling identifies regions in a genome that are enriched with aligned reads. INsPeCT uses the R package BayesPeak and peak-calling by coverage value for this purpose.[51] We input sorted BAM files to these methods, which return transcription factor-bounded regions in CSV format to the user.

Peak annotation, filtering, and extracting peak sequence regions. The R package ChiPpeakAnno[8] is employed to retrieve the gene location, distance relative to the corresponding transcription start site (TSS), and further annotations. We currently use the latest human genome assembly, GRCh37, for this purpose. At the end of this step we provide peaks

annotated with chromosomal peak location, strand, feature and TSS information, gene symbol, and Ensembl and Entrez gene IDs in a CSV file. We also provide plots with distance to TSS for visualization. Using the annotated peak data we perform GO enrichment analysis using the hypergeometric test with a BH-adjusted *P*-value cutoff set to 0.05. The output of GO enrichment analysis is provided as a CSV file. We provide functionality to extract the promoter sequences in FASTA format for a user-specified region that can be visualized in any standard genome browser.

Normalization and differential binding analysis. Researchers can perform differential binding analysis of the peaks identified in two different conditions, eg in a case/control experimental design. We perform reads per kilobase of sequence range per million mapped reads (RPKM) normalization and quality control plots for samples. Differential binding analysis can be carried out using two approaches: via the DESeq package, or using the edgeR package .[52,53] If all these tools are chosen, we compute overlaps of differential peaks across methods and provide consensus peaks for user-specified fold-change, number of peaks, and a Venn plot for visual inspection of the result.

Motif analyses. Motif discovery is of obvious relevance in ChIP-seq analysis. INsPeCT integrates the widely used MEME Suite of tools[18] for motif discovery, comparison, and analysis. We use MEME-ChIP,[54] which was specifically designed for analysis of ChIP-seq data. MEME-ChIP performs different motif analyses on the input data and includes the MEME,[55] TOMTOM,[56] SPAMO,[57] DREME,[58] CENTRIMO,[59] and AME[60] tools. An interactive HTML file is provided that summarizes the results and provides links to the results for each program. It also displays interactive plots for visual inspection.

Redirecting interesting genes to genelist analysis framework. An interesting gene list identified in ChIP-seq analysis, for instance differential peaks, can be redirected to the gene-list analysis framework (Fig. 2D) for additional functional analysis.

C. RNA-seq data. The RNA-seq analysis framework of INsPeCT (Fig. 2C and D) provides a complete analysis workflow for raw or processed data analysis. Below we describe the critical steps and functions. For the different options selected, INsPeCT automatically prepares the input data for downstream analyses.

File-format conversions and data processing. INsPeCT provides a data-upload functionality for the raw sequence read formats FASTQ and SRA, and for mapped and aligned reads in the SAM or BAM formats. We also provide functionality for automated online data import from the ArrayExpress database using the ArrayExpressHTS package.[61]

Alignment, annotation, normalization, and dispersion estimation. We provide three aligner options, Tophat,[62] Bowtie,[13] and Bwa,[12] for aligning reads against a reference genome or transcriptome. A read count table in CSV format is provided as output. For annotating read count data we use Bioconductor's human annotation database,[63] and apply the Trimmed Mean of M component technique[64] for normalization of read counts. We provide multi-dimensional scaling plot for visual inspection. We use edgeR[53] to estimate the overall dataset dispersion to detect the overall biological variability, followed by gene-wise dispersion estimation for detecting possible trend in average count size.

Differential gene expression analysis. To detect differentially expressed genes we implement the edgeR package.[53] The BH-adjusted *P*-value threshold for significance is 0.05. Different CSV files are produced giving results for coefficients, fitted values, gene annotations, and differential expression. List of significant genes with read counts are produced for further analyses. A gene list can also be redirected to the gene-list analysis framework (Fig. 2D) for functional analysis.

D. Gene-list analysis framework. The gene-list analysis framework (Fig. 2E) has been developed to provide functional analysis for researchers who may not have raw data for analysis but have previously identified interesting genes using another package or approach. Below we summarize some of the analysis options. For the different options selected, INsPeCT automatically prepares the input data for further analyses.

Promoter analysis. Automated retrieval of genomic sequences, annotation of promoter sequences, detection of CpG islands, and sequence analysis can be performed using Biomart resources[65] and the MEME Suite of tools.[18]

Motif discovery and analysis. With MEME Suite,[18] a researcher can discover motifs using MEME and DREME, search sequence databases with motifs using MAST,[66] compare a motif to all motifs in a database using TOMTOM, associate motifs with GO terms via their putative target genes using GOMO, or analyze motif enrichment using CentriMo.

Druggability analysis. Proteins that are the targets of current Food and Drug Administration (FDA)-approved anticancer drugs can be identified using a druggability analysis function developed using the Cancer Resource[67] database.

Enrichment analysis. Gene set enrichment analysis (GSEA) and GO enrichment analysis are performed using the GSEA and DAVID tools through the GSEA API[46] and DAVID-WS,[68] respectively. Gene enrichment disequilibrium analysis are performed using the R package EDanalysis.[69]

Transcription factor binding site (TFBS) over-representation analysis. The identification of over-represented single or combinations of TFBSs in sets of co-expressed genes can be performed using oPOSSUM.[70]

E. Automated workflows. Network inference can be a powerful tool in understanding how interactions are disrupted and rewired and identifying novel regulatory interactions and broader systemic disruptions in key oncogenic processes. We provide users with two automated workflows for module-based GRN inference and analysis to understand genetic architecture and underlying biology in a given system: WGCNA and

a novel method that we call RMaNI. We briefly summarize these workflows below.

WGCNA. This workflow is based on a general framework for WGCNA available as an R package.[71,72] It finds modules of highly correlated genes across microarray samples, associated with the external sample traits. We provide functions for automated network construction, module detection, gene selection, calculations of topological properties, and data visualization, and export the network in Cytoscape- and VisAnt-compatible formats.[73,74] This workflow takes processed expression data and associated patient information as input, and provides several output files in CSV format and graphics in portable network graphics (PNG) format. This workflow is suitable for users who have access to microarray expression data and clinical metadata such as survival information. Importantly, this approach does not require expression data for normal controls.

RMaNI. RMaNI is a novel analytical workflow we developed for the inference and analysis of cancer-subtype specific modules.[75] It implements the LeMoNe algorithm[76] for model-based co-clustering of expression data, and RIF[29] to identify relevant regulatory factors. LeMoNe uses a Gibbs sampling procedure to iteratively update the cluster assignment of both genes and conditions. It takes processed expression data as input, and provides several output files in CSV format and graphics in PNG format. This workflow is suitable for users who have access to microarray expression data for normal and multiple cancer subtypes. One significant advantage of LeMoNe over WGCNA is that it uses a model-based approach for clustering genes, and while selecting thresholds does not assume that networks necessarily have a scale-free topology.

Conclusion

INsPeCT is an innovative framework that provides an easy-to-use interface to a comprehensive, integrated suite of tools for rapid in-silico analysis of microarray, ChIP-seq and RNA-seq data, and/or lists of genes. It also provides access to the novel analytical framework RMaNI, and to the widely used WGCNA tool for inference and analysis of transcriptional regulatory networks using microarray data. Our web server makes available a set of tools and analytical workflows that would otherwise be challenging for non-expert users to install and apply. In future, we will integrate more tools and workflows to meet the distinct needs of researchers confronting the complexity of cancer transcriptomics.

List of Abbreviations Used

GEO: gene expression omnibus
GRN: gene regulatory network
RMaNI: regulatory module network inference
JT: Jonckheere—Terpstra
RIF: regulatory impact factors
TF: transcription factor
TFBS: transcription factor binding site
RN: relevance networks
MRNET: minimum redundancy/maximum relevance networks
CLR: context likelihood relatedness
ARACNE: the algorithm for the reconstruction of accurate cellular networks
PCIT: partial correlation and information theory
WGCNA: weighted gene co-expression network analysis
LeMoNe: learning module networks
GENIE: gene network inference with ensemble of trees
GO: gene ontology
GSEA: gene set enrichment analysis
CSV: comma-separated values
PNG: portable network graphics
RPKM: reads per kilobase of sequence range per million mapped reads
FDA: Food and Drug Administration

Author Contributions

PBM developed and wrote the code for INsPeCT, and composed the manuscript. SRM, MJD, AR, and MAR advised on the design and features of INsPeCT, provided overall scientific and technical guidance, and assisted with the creation of the manuscript. All authors reviewed and approved of the final manuscript.

Acknowledgements

We thank Dr Timothy Bailey for valuable advice on implementation of the MEME Suite of tools and CpG detection program; Mr Gavin Graham, Dr Gerald Hartig, and Mr Alex Varlokov from Bioinformatics Resource Australia-EMBL for high-performance computing and web-development support; Dr Sriganesh Srihari and Dr Daniel Hurley for helpful discussions; Dr Richard Newton for help with Rwui; and the R/Bioconductor research community, who have made their programs and source codes publicly available. Computational resources were provided by National Computational Infrastructure Specialised Facility in Bioinformatics.

DISCLOSURES AND ETHICS

As a requirement of publication the authors have provided signed confirmation of their compliance with ethical and legal obligations including but not limited to compliance with ICMJE authorship and competing interests guidelines, that the article is neither under consideration for publication nor published elsewhere, of their compliance with legal and ethical guidelines concerning human and animal research participants (if applicable), and that permission has been obtained for reproduction of any copyrighted material. This article was subject to blind, independent, expert peer review. The reviewers reported no competing interests.

REFERENCES

1. [http://cancergenome.nih.gov/].
2. [http://icgc.org/].
3. Kodama Y, Shumway M, Leinonen R. The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.* 2012;40(Database issue):D54–6.

4. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002;30(1): 207–10.
5. Brazma A, Parkinson H, Sarkans U, *et al*. ArrayExpress – a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res*. 2003;31(1): 68–71.
6. Parkinson H, Kapushesky M, Kolesnikov N, *et al*. ArrayExpress update – from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res*. 2009;37(Database issue):D868–72.
7. Green ED, Guyer MS. Charting a course for genomic medicine from base pairs to bedside. *Nature*. 2011;470(7333):204–13.
8. Zhu LJ, Gazin C, Lawson ND, et al. ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics*. 2010;11:237.
9. Zambelli F, Prazzoli GM, Pesole G, Pavesi G. Cscan: finding common regulators of a set of genes by using a collection of genome-wide ChIP-seq datasets. *Nucleic Acids Res*. 2012;40(W1):W510–15.
10. Medina I, De Maria A, Bleda M, et al. VARIANT: Command Line, Web service and Web interface for fast and accurate functional characterization of variants found by Next-Generation Sequencing. *Nucleic Acids Res*. 2012;40(W1): W54–8.
11. Boeva V, Lermine A, Barette C, Guillouf C, Barillot E. Nebula – a web- server for advanced ChIP-seq data analysis. *Bioinformatics*. 2012;28(19):2517–9.
12. Li H, Durbin R. Fast and accurate long-read alignment with Burrows—Wheeler transform. *Bioinformatics*. 2010;26(5):589–95.
13. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9(4):357–9.
14. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841–2.
15. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–9.
16. Medina I, Carbonell J, Pulido L, *et al*. Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. *Nucleic Acids Res*. 2010;38(Web Server issue):W210–3.
17. Linderman GC, Chance MR, Bebek G. MAGNET: MicroArray Gene expression and Network Evaluation Toolkit. *Nucleic Acids Res*. 2012;40(W1):W152–6.
18. Bailey TL, Boden M, Buske FA, et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res*. 2009;37(Web Server issue):W202–8.
19. De Smet R, Marchal K. Advantages and limitations of current network inference methods. *Nat Rev Microbiol*. 2010;8(10):717–29.
20. Michoel T, Maere S, Bonnet E, *et al*: Validating module network learning algorithms using simulated data. *BMC Bioinformatics*. 2007;8(suppl 2):S5.
21. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*. 2013;14:91.
22. R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2012.
23. Gentleman RC, Carey VJ, Bates DM, *et al*. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*. 2004;5(10): R80.
24. Newton R, Wernisch L. Rwui: a web application to create user friendly web interfaces for R scripts. *R News*. 2007;7(2):32–5.
25. Fan J, Ren Y. Statistical analysis of DNA microarray data in cancer research. *Clin Cancer Res*. 2006;12:4469–73.
26. Gordon S. Limma: linear models for microarray data. In: Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W, eds. *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. New York: Springer; 2005:397–420.
27. Schwende H. siggenes. *Multiple Testing Using Sam and Efron's Empirical Bayes Approaches*. R Package Version 1310. Published 2012.
28. Broberg P. *SAGx: Statistical Analysis of the GeneChip*. 2008. R Package Version 1310.
29. Reverter-Gomez A, Hudson NJ, Nagaraj SH, Perez-Enciso M, Dalrymple BP. Regulatory impact factors: unraveling the transcriptional regulation of complex traits from expression data. *Bioinformatics*. 2010;26(7):896–904.
30. Butte AJ, Kohane IS. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput*. 2000: 418–29.
31. Margolin AA, Nemenman I, Basso K, et al. ARACNE: an algorithm for the reconstruction of s in a mammalian cellular context. *BMC Bioinformatics*. 2006;7(suppl 1):S7.
32. Faith JJ, Hayete B, Thaden JT, et al. Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol*. 2007;5(1):e8.
33. Meyer PE, Kontos K, Lafitte F, Bontempi G. Information-theoretic inference of large transcriptional regulatory networks. *EURASIP J Bioinform Syst Biol*. 2007:79879.
34. Reverter A, Chan EKF: Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks. *Bioinformatics*. 2008;24(21):2491–7.
35. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring regulatory networks from expression data using tree-based methods. *PloS One*. 2010;5(9):e12776.
36. Meyer P, Lafitte F, Bontempi G. minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics*. 2008;9(1):461.
37. Watson-Haigh NS, Kadarmideen HN, Reverter A. PCIT: an R package for weighted gene co-expression networks based on partial correlation and information theory approaches. *Bioinformatics*. 2010;26(3):411–3.
38. Matys V, Fricke E, Geffers R, *et al*. TRANSFAC(R): transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*. 2003;31(1):374–8.
39. Schaefer U, Schmeier S, Bajic VB. TcoF-DB: dragon database for human transcription co-factors and transcription factor interacting proteins. *Nucleic Acids Res*. 2011;39(Database issue):D106–10.
40. Madhamshettiwar PB, Maetschke SR, Davis MJ, Reverter A, Ragan MA. Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets. *Genome Med*. 2012;4(5):41.
41. Fang C, Weiliang Q, Ruben HZ, Ross L, Xiaogang W. clues: an R package for nonparametric clustering based on local shrinking. *J Stat Software*. 2010;33(4): 1–16.
42. Fraley C, Raftery AE. *MCLUST Version 3: An R Package for Normal Mixture Modeling and Model-Based Clustering*. Seattle, WA, USA: Department of Statistics, University of Washington; 2006.
43. Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K. *cluster: Cluster Analysis Basics and Extensions*. R Package Version 1143. Published 2012.
44. Tarca AL, Draghici S, Khatri P, et al: a novel signaling pathway impact analysis. *Bioinformatics*. 2009;25(1):75–82.
45. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 2000;28:27–30.
46. Subramanian A, Tamayo P, Mootha VK, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545–50.
47. Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ. Discovering statistically significant pathways in expression profiling studies. *Proc Natl Acad Sci U S A*. 2005;102(38):13544–9.
48. Wang C, Taciroglu A, Maetschke SR, Nelson CC, Ragan MA, Davis MJ. mCOPA: analysis of heterogeneous features in cancer expression data. *J Clin Bioinforma*. 2012;2(1):22.
49. [http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=show&f=software&m=software&s=software].
50. [http://www.bioinformatics.babraham.ac.uk/projects/fastqc/].
51. Spyrou C, Stark R, Lynch AG, Tavare S, BayesPeak. Bayesian analysis of ChIP-seq data. *BMC Bioinformatics*. 2009;10:299.
52. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol*. 2010;11(10):R106.
53. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26(1):139–40.
54. Machanick P, Bailey TL. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*. 2011;27(12):1696–1697.
55. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*. 1994;2: 28–36.
56. Gupta S, Stamatoyannopoulos J, Bailey T, Noble W. Quantifying similarity between motifs. *Genome Biol*. 2007;8(2):R24.
57. Whitington T, Frith MC, Johnson J, Bailey TL. Inferring transcription factor complexes from ChIP-seq data. *Nucleic Acids Res*. 2011;39(15):e98.
58. Bailey TL. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*. 2011;27(12):1653–1659.
59. Bailey TL, Machanick P. Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res*. 2012;40(17):e128.
60. McLeay RC, Bailey TL. Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics*. 2010;11:165.
61. Goncalves A, Tikhonov A, Brazma A, Kapushesky M. A pipeline for RNA-seq data processing and quality assessment. *Bioinformatics*. 2011;27(6): 867–9.
62. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009;10(3):R25.
63. Carlson M. *org.Hs.eg.db: Genome Wide Annotation for Human*. R Package Version 2.8.0.
64. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*. 2010;11(3):R25.
65. Durinck S, Moreau Y, Kasprzyk A, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*. 2005;21(16):3439–40.
66. Bailey TL, Gribskov M. Combining evidence using *P*-values: application to sequence homology searches. *Bioinformatics*. 1998;14(1):48–54.
67. Ahmed J, Meinel T, Dunkel M, et al. CancerResource: a comprehensive database of cancer-relevant proteins and compound interactions supported by experimental knowledge. *Nucleic Acids Res*. 2011;39(Database issue):D960–7.

68. Jiao X, Sherman BT, Huang da W, et al. DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics*. 2012;28(13):1805–6.
69. Jiang Y. *EDanalysis: A R Package For Gene Enrichment Disequilibrium Analysis*. R Package Version 101. Published 2012.
70. Ho Sui SJ, Fulton DL, Arenillas DJ, Kwon AT, Wasserman WW. oPOSSUM: integrated tools for analysis of regulatory motif over-representation. *Nucleic Acids Res*. 2007;35(Web Server issue):W245–52.
71. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*. 2008;9(1):559.
72. Zhang B, Horvath S. A general framework for weighted gene co- expression network analysis. *Stat Appl Genet Mol Biol*. 2005;4:Article17.
73. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13(11):2498–504.
74. Hu Z, Mellor J, Wu J, Yamada T, Holloway D, Delisi C. VisANT: data-integrating visual framework for biological networks and modules. *Nucleic Acids Res*. 2005;33(Web Server issue):W352–7.
75. Madhamshettiwar PB, Maetschke SR, Davis MJ, Reverter A, Ragan MA. RMaNI: regulatory module network inference framework. *BMC Bioinformatics*. 2013;14(suppl 12):S14.
76. Joshi A, Van de Peer Y, Michoel T. Analysis of a Gibbs sampler method for model-based clustering of gene expression data. *Bioinformatics*. 2008;24(2):176–83.