iDEP: An integrated web application for differential expression and pathway analysis

Steven Xijin Ge
Department of Mathematics and Statistics, South Dakota State University, Brookings, SD 57007
gexijin@gmail.com

Abstract: iDEP (integrated Differential Expression and Pathway analysis) is a web application that reads in gene expression data from DNA microarray or RNA-Seq and performs exploratory data analysis (EDA), differential expression, and pathway analysis. The key idea of iDEP is to make many powerful R/Bioconductor packages easily accessible by wrapping them under a graphical interface, alongside annotation databases. For EDA, it performs hierarchical clustering, k-means clustering, and principal component analysis (PCA). iDEP detects differentially expressed genes using the limma and DESeq2 packages. For a group of co-expressed genes, it identifies enriched gene ontology (GO) terms as well as transcription factor binding motifs in promoters. Pathway analysis can be performed using several packages like GAGE, GSEA, PGSEA, or ReactomePA. iDEP can also detect chromosomal gain or loss using the PREDA package. iDEP uses annotation of 69 metazoa and 44 plant genomes in Ensembl for ID mapping and functional categorization. Pathway information was also compiled from databases like KEGG, Reactome, MSigDB, GSKB, and araPath. As an example, we extensively analyzed an RNA-Seq dataset involving siRNA-mediated Hoxa1 knockdown in lung fibroblasts, and identified the down-regulation of cell-cycle genes, in agreement with previous findings. Our analyses also reveal the possible roles of E2F1 and its target genes, including microRNAs, in blocking G1/S transition, and the upregulation of genes related to cytokines, lysosome, neuronal parts. By integrating many R and Bioconductor packages with comprehensive annotation databases, iDEP (http://ge-lab.org/idep/) enables users to conduct in-depth bioinformatics analysis of transcriptomic data through a graphical interface.

Background

With low cost and high resolution, RNA-Seq is becoming a widely-used tool for transcriptional profiling. While library construction and sequencing can be carried out routinely following protocols, analysis and interpretation of the resultant data remain a challenge, especially for biologists without bioinformatics training. Based on the powerful statistical computing language R, the Bioconductor[1] platform enables the sharing of high-quality, modularized R code, together with annotation databases, as thousands of software packages. These packages greatly simplify and streamline many common analysis workflows. It is one of the most important developments in bioinformatics in recent years. These R packages, however, are inaccessible to researchers without programming experience. To mitigate this issue, we aim to develop a user-friendly web-based application that encompasses many useful R and Bioconductor packages as well as annotation databases.

RNA-Seq data analysis often starts with pre-processing, mapping and summarizing of short sequencing reads. We assume that this step has been completed, using either the traditional Tuxedo Suite[2,3], or other alternatives such as the faster, alignment-free quantification methods[4,5]. Also available are stand-alone software like GenePattern[6], as well as web-based solutions including Galaxy[7] and iPlant/CyVerse[8].

After read mapping, we often obtain a matrix of gene-level read counts or normalized expression levels. For such data, many R and Bioconductor packages have been developed to conduct exploratory data analysis (EDA), identification of differentially expressed genes (DEGs), and pathway analysis. Using

Shiny[9], a framework that can easily turn R code into interactive web applications, several tools have been developed to enable researchers to use these R packages via a graphical user interface (GUI). START App (Shiny Transcriptome Analysis Resource Tool) is a Shiny app that performs hierarchical clustering, principal component analysis (PCA), gene level boxplot, and differential gene expression[10]. Another similar tool, Degust[11] can perform differential expression analysis using EdgeR[12] or limma-voom[13], and interactively plot the results. Other tools include Sleuth[14] and ShinyNGS[15]. Non-Shiny applications were also developed to take advantage of the R code base. This includes DEIVA [16] and VisRseq[17]. Beyond differential expression, several tools incorporate some capacity of pathway analysis. For quantified expression data, ASAP (Automated Single-cell Analysis Pipeline)[18] can carry out normalization, filtering, clustering, and enrichment analysis based on Gene Ontology (GO)[19] and KEGG[20] databases. With EXPath Tool[21], users can perform pathway search, GO enrichment and co-expression analysis. The development of these tools in the last few years facilitate the interpretation of large genomics data.

Inspired by these efforts, we seek to further enhance the capacity of such systems by including (1) automatic gene ID conversion with broad coverage, (2) multiple methods for pathway analysis and visualization, and (3) comprehensive pathway database. Leveraging the massive amount of gene annotation information in Ensembl[22,23] and Ensembl Plants[24], as well as pathway databases compiled by our group[25,26] and others[20,27,28], we were able to build a large database to support in-depth analyses of expression data in over a hundred species. With such data and the inclusion of many R packages for pathway analysis and visualization (see flow chart in Figure 1), iDEP (integrated Differential Expression and Pathway analysis) enables users to easily formulate new hypotheses from expression datasets. To demonstrate the use of iDEP, we analyzed an example dataset and generated all the tables and figures (except Figure 1) in this paper.

Results

iDEP design

Our goal is to build an easy-to-use Shiny app that integrates many commonly-used R/Bioconductor packages with comprehensive annotation databases. To reduce complexity, we focus on the analysis of gene-level expression matrices obtained from read mapping using other tools. We try to develop an intuitive, graphical, and robust tool so that researchers without bioinformatics experience can routinely and quickly translate expression data into actionable insights and testable hypotheses. We also want to make an open system where users can download intermediate results so that other tools can be used, or upload custom pathway databases for unannotated species. For experienced bioinformaticians, it can serve as a tool for preliminary analysis as it circumcises the need for many tedious tasks such as converting gene IDs and downloading software packages and annotations. As shown in the flowchart in Figure 1, normalized expression and RNA-Seq read counts are handled by two analysis workflows. Both involve a 4-stage process: (1) pre-processing, (2) EDA, (3) differential expression, and (4) pathway analysis and visualization.

To enable gene ID conversion, we massively downloaded all available gene ID mappings for 69 metazoa genomes from Ensembl[22,23] and 44 plant genomes from Ensembl Plants[24]. See Table S1 in Supplementary File 1 for a list of these species. The final mapping table consists of 71,766,733 rows, linking 1645 types of IDs (Table S2 in Supplementary File 1) with Ensembl IDs in 111 species. For example, 64 types of human gene IDs can be converted to Ensembl gene IDs, which are used internally to identify genes. Besides common ID types like gene symbol, Entrez, Refseq, UCSC, UniGene, and Interpro IDs, the 64 kinds of human gene IDs also include probe IDs for 31 popular DNA microarray platforms, making it possible to re-analyze thousands of microarray datasets available at public repositories.

In the pre-process stage, user's gene IDs are first compared to all gene IDs in the database for all organisms. This enables automatic ID conversion and simultaneous species identification. Genes expressed at very low levels are removed and data are transformed as needed using one of several methods. iDEP

enforces log-transformation, when skewed distribution is detected. This type of mechanisms can help avoid issues in downstream analyses that assumes approximate normality. The pre-process stage also generates diagnostic and summary plots to guide users to make their choices. We also try to minimize unnecessary mouse clicks by automatically performing related analyses and show multiple results on one page.

EDA enables the users to explore variations and patterns in the dataset as a whole[29]. The main methods include hierarchical clustering with heatmap, k-means clustering, and PCA. Enrichment analysis of genes derived from k-means clustering is conducted to gain insights on the functions of co-expressed genes. Initial attempts of pathway analysis are carried out using the PCA loadings on each gene. This can tell us the biological processes underlying each direction of expression change defined by the principal components.

Differential expression analysis relies on two Bioconductor packages, *limma*[30] and DESeq2[31]. These two packages can meet the needs for most studies, including those involving multiple biological samples and factorial design. Normalized expression data is analyzed using *limma*. Read counts data can be analyzed using three methods, namely limma-trend[13], limma-voom[13,32], and DESeq2. Other methods such as edgeR [12] may be incorporated in the future.

Fold-change values for all genes returned by *limma* or DESeq2 are used in pathway analysis using GSEA[33], PAGE[34,35], GAGE[36] or ReactomePA[37]. Taking advantage of centralized annotation databases at Ensembl[22,23] and Ensembl Plants[24], we downloaded not only GO functional categorizations, but also promoter sequences for defining transcription factor (TF) binding motifs for most of the 111 species. Metabolic pathways were downloaded directly from KEGG [20] for 67 species (Table S1 in Supplementary File 1). In addition, there are many species-specific pathway knowledgebases, such as Reactome[28,37] and MSigDB[27] for human, GSKB for mouse[25], and araPath for Arabidopsis[26]. These databases contain diverse types of gene sets, ranging from TF and microRNA target genes to manually curated lists of published DEGs. We downloaded and converted these databases to enable in-depth analysis of expression data from different perspectives. Also, we incorporated Pathview package[38] to show gene expression on KEGG pathway diagrams. Based on genes' chromosomal location from Ensembl, we visualize fold-changes on the chromosomes and also use the PREDA package[39] to detect chromosomal regions enriched with DEGs.

Example data

We used iDEP to analyze a dataset related to the effect of Hoxa1 knockdown by short interfering RNA (siRNA) in human lung fibroblasts[3]. This RNA-Seq dataset was used as example data for the development of Cuffdiff2[3]. Turner analyzed it in a tutorial[40] for RNA-Seq data analysis using gene-level counts generated by Sailfish[41]. This read count data (Supplementary File 2) is used in our analysis and is also built into iDEP as demo data.

Pre-processing and EDA

Using the "Click here to load demo data" button, we can load the data into iDEP, which correctly recognized Homo sapiens as the most likely species with the most genes IDs matched. After ID conversion and the default filter, 13,783 genes left. A bar plot of total read counts per library is generated (Figure 2A), showing some variation in library size. We choose the regularized log (rlog) transformation implemented in the DESeq2 package. The transformed data can be downloaded (Supplementary Table S3); its distribution is shown in Figure 2B-C. Variation among technical replicates is relatively small, even among lowly expressed genes (Figure 2D).

Data transformation has large effects on many lowly expressed genes. Figure 3 illustrates the effect of various methods using data from [42]. For started log (log(x+c)), as the pseudo count c increases, it becomes increasingly aggressive in rounding up smaller values. The rlog method, which is much slower, preserves

smaller numbers while reducing variability. Variance stabilizing transform (VST) [43] converts all smaller expression scores to above 5, similar to started log with a large pseudo count.

The "Barplot for one or more genes" button enables users to examine the expression of a specific gene or gene family. Using “Hoxa” as a keyword, we obtain Figure 4A, which shows that Hoxa1 expression level is reduced, but not abolished, in response to siRNA-mediated knockdown of Hoxa1. Noticeably, expression of other family members, especially Hoxa2, 4, and 5, are also decreased. As these genes have similar mRNA sequences, it is unclear whether this is caused by off-target effects of the siRNA, or ambiguous mapping of RNA-Seq reads. Figure 4B, obtained by using “E2F” as keyword, shows the decreased expression of E2F1 and other genes, which will be discussed later.

For clustering analysis, we rank genes by their standard deviation across all samples. The result of hierarchical clustering of the top 1000 genes is shown in Figure 5A. The plot suggests that the expression pattern in the two types of cells are clearly distinct, with hundreds of genes differentially expressed. Variations among technical replicates are small. An interactive version based on the *Plotly* enables users to zoom in to see gene names and generate plots like Figure 5B. A correlation matrix containing Pearson’s correlation coefficient (Figure 5C) again indicates minimal variation among technical replicates.

Next, we run k-means clustering of the top 2000 genes ranked by standard deviation. By default, the genes are divided into 4 clusters. The optimal number of clusters can be determined by plotting of within-groups sum of squares by clicking on a button labeled “How many clusters?”. Figure 6A suggests that increasing *k* from 2 to 3 does not substantially improve the modeling of data. After choosing *k*=2, we get Figure 6B. Details of these clusters can be downloaded (Supplementary Table S4). Enrichment of GO terms in these clusters is automatically generated (Table 1). The 1102 genes in Group A are upregulated by Hoxa1 knockdown and is overrepresented with genes related to cell cycle. The 898 genes in Group B genes are suppressed by the knockdown and enriched with genes associated with response to endogenous stimulus, cell migration, and locomotion.

Significant enrichments are observed for TCFL5, E2F4, SP2 and MBD2 binding motifs in the 300bp upstream promoters of groups A genes and to a less extend group B genes. Note that the binding motifs of a family of TFs are often similar. Therefore the specific family member reported might not be reliable. E2F factors are regulators of cell cycle [44]. There are some reports that SP2 also modulate cell cycle [45]. The effect of Hoxa1 knockdown on cell cycle was reported and experimentally confirmed in the original study [3]. Cell cycle analysis revealed that loss of Hoxa1 leads to a block in G_1 phase [3]. E2F1 promotes G_1/S transition [46] by regulation many genes, including itself. Therefore, our results not only agree with the original study, but also provide possible gene regulatory mechanisms.

PCA plot using the first and second principal components is shown in Figure 7. There is a clear difference between the Hoxa1 knockdown and the control samples, along the first principal component, which explains most (93%) of the variance. Plot using multidimensional scaling (MDS) also show a similar distribution of the samples. We can choose to conduct pathway analysis using PGSEA [34,35] by treating the loadings of the principal components on the genes as expression data. The first two components are related to cell cycle regulation, confirming our results based on k-means clustering.

Differentially Expressed Genes (DEGs)

With the default DESeq2 method, we identified 907 upregulated and 1097 downregulated genes (see Supplementary Table S5) using a threshold of FDR < 0.1 and fold-change >2. The Venn diagram is shown in Figure 8A. Venn diagrams are more informative when we have DEGs from multiple comparisons. The volcano plot (Figure 8B) and the scatter plot (Figure 8C) show that Hoxa1 knockdown leads to a massive transcriptomic response with hundreds of significant genes. Plotly-based interactive versions of these plots

are available, where users can zoom in and mouse over to see individual genes (Figure 8D). We also get another heatmap that visualizes the expression pattern of the selected DEGs (data not shown). The top genes ranked by the absolute values of fold-changes are also given as a table (Supplementary Table S6). A quick scan at the lists tell us that Hoxa1 knockdown induces several cytokines (IL1B, IL24).

The GO terms enriched in DEGs are shown in Table 3. Upregulated (Hoxa1KN > control) genes are related to regulation of cell proliferation, locomotion, and response to endogenous stimulus. This is perhaps the cell's response to injected siRNAs. The downregulated genes (Hoxa1KN < control) are enriched with cell cycle-related genes (FDR < 2.6×10^{-47}), consistent with the original study. As shown in Table 4, the promoters of differentially expressed genes are overrepresented with many G-rich motifs bound by E2F and other factors such as TCFL5 and SP2.

Besides GO, there are many other types of gene sets for enrichment analysis. The Motif gene sets from MSigDB are derived from [47], and contain sets of genes sharing TF binding motifs in gene promoters and microRNA target motifs in 3' untranslated regions (UTRs). Using this gene set, we again detect the enrichment of E2F motifs in promoters (Table 5). Table 5 also suggest an overrepresentation of a "GCACTTT" motif in 3' UTRs of upregulated genes. This motif is targeted by several microRNAs such as miR-17-5P, miR-20a, miR-106a. Cloonan *et al.* showed that miR-17-5P targets more than 20 genes involved in the G_1/S transition [48]. Trompeter *et al.* provided evidence that miR-17, miR-20a, and miR-106b enhance the activities of E2F factors to influence G_1/S transition [49]. miR-106b resides, along the sense direction, in the intron of Mcm7 gene, an E2F1 target gene that is also downregulated by Hoxa1 knockdown (see Figure 11A). Petrocca *et al.* showed that E2F1 regulates miR-106b, which can conversely control E2F1 expression [50]. Thus, it is possible that reduced E2F1 expression lead to decreases of these microRNAs, which causes the increases in the expression of their target genes. Leveraging the comprehensive pathway databases, iDEP enables researchers to develop new hypotheses that could be further investigated.

Choosing GO cellular component, we find that Hoxa1 knockdown suppresses genes that code for spindle, chromosomal parts, and microtube (Table 6). As Hoxa1 knockdown blocks G_1/S transition [3], smaller number of cells are in the S (synthesis) phase, leading to the reduction of proteins related to spindle and chromosomal parts. Hoxa1 knockdown also induces genes related to plasma membrane, neurons and synapses (Table 6). This unexpected result is consistent with the finding that Hoxa1 mutations affect neuronal differentiation in neuroblastoma cells [51] and can lead to neuronal defects in the hindbrain of mutant mice [52]. Polymorphisms of this gene are associated with cerebellar volume in humans [53]. Even in lung fibroblasts, Hoxa1 knockdown induce neuron related genes.

Choosing KEGG pathway, we confirm the overrepresentation of cell cycle-related genes in downregulated genes. For up-regulated genes, we detect cytokine-cytokine receptor interaction (CCRI) pathway (FDR<1.3×10^{-10}). "MSigDB.Curated" gene sets contain pathways from pathway databases as well as published lists of DEGs from previous expression studies. The most significant (FDR<10^{-100}) results were the overlap of downregulated genes with many published DEGs. By searching for these gene sets by name (i.e. CHANG_CYCLING_GENES) via Google, we can get direct links to MSigDB. At least one of top sets is related to cell cycle [54]. As suggested by a meta-analysis of published gene lists [55], cell cycle related expression signature are frequently shared by diverse cellular perturbations. Indeed, many conditions may affect cell cycle [56]. We uncover similarity of our expression signature with previously published ones.

MSigDB also contains other types of gene sets such as cytogenetic band, computational gene clusters, immune related genes, and cancer-related genes. Our group has also developed a similar database for mouse called GSKB [25], which is also included in iDEP. GSKB contain multiple sources of TF and microRNA

target genes for identification of gene regulatory mechanisms. Beyond GO, such comprehensive knowledgebases are essential for in-depth interpretation of expression data.

Pathway analysis

Instead of using lists of DEGs that are sensitive to arbitrary cutoffs, pathway analysis can use fold-changes of all genes. At the "Pathways" tab, we use the default GAGE [36] as method and KEGG as gene sets. The result in Table 7 is similar to those from a previous analysis [40] and also agrees with our enrichment analysis based on DEGs. For each of the significant KEGG pathways, we can view the fold-changes of related genes on a diagram using Pathview package [38]. Many cell cycle genes are marked as green on Figure 9, indicating reduced expression in Hoxa1-knockdown samples. We also detected upregulation of genes related to CCRI, arthritis, and lysosome. Not detected using DEGs, lysosome-related genes are mostly upregulated (Figure 10). It is possible that injected siRNAs triggers cytokines, as IL1B, IL24 are highly induced, and is subsequently degraded in the lysosome.

Changing pathway gene sets, we can gain further insights. Using MSigDB.Motif gene sets, we can verify the enrichment of E2F binding motifs. For non-KEGG gene sets, heatmaps are created to show the expression of genes in significant gene sets. Figure 11A shows part of such a plot, highlighting genes that share the "SGCGSSAAA" motif bound by E2F1. Note that E2F1 gene itself is included in the figure, as it binds to its own promoter and form a positive feedback loop [46]. Another gene, Mcm7, hosts miR-106 in its intron, a microRNA that can target E2F1 transcripts, forming a negative feedback loop [50].The downloaded expression data indicates that E2F1 is downregulated by more than 3-fold in Hoxa1 knockdown samples (also see Figure 4B). Downregulation of E2F and downstream genes, including microRNAs, may be part of the transcription program that blocks G_1/S transition.

Users can use many combinations of methods and gene sets to conduct pathway analysis. For example, using PGSEA on KEGG pathways yielded Figure 11B, again confirming previous results on suppressed cell cycle genes and induced lysosome and CCRI related genes. Using the MSigDB.Motif gene sets, we can also confirm the E2F1 binding motifs (Figure 11C). The most highly activated gene sets are related to miR-17-5p, miR-20a, miR106a and so on (Figure 11C), which agrees with our analysis using just gene lists.

Some pathways can be attenuated by upregulating some of its genes while downregulating others. To detect such pathways, we can use the absolute values of fold changes in pathway analysis. This is achieved by checking the box labeled "Use absolute values of fold changes for GSEA and GAGE." We detected the downregulation of genes involved in ribosomes and neurodegenerative diseases, such as Huntington's disease and Parkinson's disease. This is again consistent with the role of this gene in neuronal differentiation [52] as discussed above.

The expression of neighboring genes can be correlated, due to many mechanisms like super-enhancers [57], 3D chromatin structure [58], or genomic gain or loss in cancer. To help users detect such correlation, we use ggplot2 [59] and Plotly to interactively visualize fold-changes on all the chromosomes (Figure 12B). The PREDA package [39] can detect statistically significant chromosomal regions with coherent expression change among neighboring genes. Figure 12B shows many such regions in response to Hoxa1 knockdown. Detailed information obtained from downloaded files (supplementary Table S7) suggests, for example, a 4.3 Mbps region on Chr.1q31 contain 6 upregulated genes (PRG4, TPR, C1orf27, PTGS2, PLA2G4A, and BRINP3).

To further validate our parameterization of PREDA, we analyzed DNA microarray data of thymus tissues from patients with down syndrome [60]. We detected large, upregulated regions on chromosome 21 (Figure 13), as expected. Even though PREDA analysis is slow and has low-resolution due to the use of gene-level

expression score, it might be useful in studies such as cancer where localized expression change on the chromosome can happen.

Hoxa1 Knockdown in mouse

As it is very easy to use iDEP to analyze published expression data, we also analyzed DNA microarray data (Supplementary File 4) generated by silencing of Hoxa1 in mouse mammary tumors using intraductal injection of siRNA lipidoid nanoparticles [61].The heatmap in Figure 14A shows that one of the control samples are more similar to Hoxa1 knockdown sample, highlighting the importance of EDA in identifying outliers. The Venn diagram in Figure 14 B summarize the overlaps among DEGs across three biological samples, which suggest that smaller changes at 16 weeks after siRNA injection and large effects at 20 weeks. Using the GSKB pathway database [25], we detect the regulation of cell cycles and the enrichment of E2F1 binding motifs in DEGs (Figure 14C&D), as expected. Surprisingly, E2F1 and other cell cycle related genes are upregulated in this study (Figure 14E), contrary to what observed in human cell lines. This can be further studied to see whether this is due to technical issues, or differences in species or tissue/cell type.

To improve reproducibility in research, we also provide an "R" tab, where users can obtain a copy of the source code as well as all user and system settings (partially shown in Figure 15). System settings include details of the software environment such as versions of R and R packages which are regularly upgraded. It is recommended that users keep a copy of such information for their record.

Discussions and conclusions

By integrating many Bioconductor packages with comprehensive annotation databases, iDEP enables users to conduct in-depth bioinformatics analysis of transcriptome data through a GUI. The Shiny platform makes it surprisingly straightforward for such an undertaking. We were able to pack many useful functionalities into iDEP, including high-quality graphics based on ggplot2 and interactive plots using Plotly. Compared with traditional web applications, Shiny has its drawbacks and limitations. With the free version, only 1 thread and 5 concurrent users are allowed. The interface is not as flexible as those developed using JavaScript. Nevertheless, we believe an integrated web application like iDEP is a valuable tool to both bench scientists and bioinformaticians.

As an example, we extensively analyzed an RNA-Seq dataset involving Hoxa1 knockdown by siRNA in lung fibroblasts, and identified the down-regulation of cell-cyle genes, in agreement with previous analyses and experimental confirmation [3]. Thanks to a comprehensive knowledgebase, our analyses also show E2F binding motifs are enriched in the promoters of downregulated genes. Thus the massive transcriptomic response induced by Hoxa1 knockdown may be partially mediated by the down-regulation of E2F factors (especially E2F1) and their target genes (Figure 11A). Furthermore, we also find limited evidence that microRNAs (miR-17-5P, miR-20a, miR-106a) might work together with E2F factors to block the G_1/S transition in response to reduced Hoxa1 expression. DEGs are also enriched with genes related to neuron parts, synapse, as well as neurodegenerative diseases. This is consistent with reports of Hoxa1's role in neuron differentiation [51-53]. Hoxa1 knockdown induces expression of genes associated with cytokine-cytokine interaction, lysosome, and cell migration, probably in response to the injected siRNAs. By hiding computational complexity and incoporating pathway databases, *iDEP empowers biologists to leverage large datasets in developing new hypothesis*.

We should be cautious in over-interpreting of results from pathway analysis. The biomedical literature is large and heterogenous [62], making it easy to rationalize and to "make a story out of any gene". As iDEP provides many different options for pathway analysis, actual results should be consistent across different settings, methods, and databases. This is evident by the repeated identification of the large effect on cell cycle

In addition to updating the annotation database derived from Ensembl every year, we plan to continue to compile pathway databases for model organisms, similar to MSigDB and GSKB. For currently unsupported species, we will consider ways to incorporate user-submitted gene annotation. Based on user request and feedback, we can also add more functions by including additional Bioconductor packages.

Methods

Input file format

Users can upload a CSV (comma-separated value) or tab-delimited text file with the first column containing gene IDs. For RNA-seq data, read count per gene is recommended, as differentially expressed genes (DEGs) can be identified based on statistical modeling of counts using DESeq2 [31]. Also accepted are normalized expression data based on FPKM (Fragments Per Kilobase of transcript per Million mapped reads), RPKM (Reads Per Kilobase of transcript per Million mapped reads), or DNA microarray data.

iDEP can automatically convert many types of common IDs to Ensembl gene IDs, which is used internally for enrichment and pathway analyses. Mappings are based on annotation of 69 metazoa species retrieved from Ensembl (version 89, accessed June 2-3, 2017) [22,23] and 42 plant species from Ensembl Plants [24]. See Table S1 in Supplementary File 1 for a list of these species. Annotation information is downloaded in batch from BioMart [63] using the biomRt Bioconductor package [64,65]. For human, 64 types of gene IDs can be mapped to Ensembl gene ID. This includes common IDs like gene symbol, Entrez ID, Refseq, UCSC, UniGene, Interpro, as well as 31 types of DNA microarray probe IDs for platforms like Affymetrix, Illumina, and Agilent. For mouse, iDEP can recognize 54 types of gene IDs. The final ID mapping table contains more than 71 million rows, linking 1756 types of IDs with Ensembl ID in 111 species. iDEP can also be used to analyze expression data in other species, as users can upload their own gene sets file.

iDEP parses column names to define sample groups. To define 3 biological samples (Control, TreatmentA, TreatmentB) with 2 replicates each, column names should be: Ctrl_1, Ctrl_2, TrtA_1, TrtA_2, TrtB_1, TrtB_2. All pair-wise comparisons (*Ctrl vs. TrtA, Ctrl vs. TrtB, TrtA vs. TrtB*) will be listed and analyzed. For factorial design, use underscore "_" to separate factors such as genetic background (wide type vs. mutant: WT vs. Mu) and experimental condition (Ctrl vs. Trt). To define an 2x2 factorial design, use column names like: WT_Ctrl_1, WT_Ctrl_2, WT_Trt_1, WT_Trt_2, Mu_Ctrl_1, Mu_Ctrl_2, Mu_Trt_1, Mu_Trt_2. A factorial design leads to more meaningful comparisons along the rows and columns of the design matrix and enables the *limma* package [30] to identify genes that respond differently to treatment in mutant lines compared to wild-type. Currently, only two factors are allowed; each can have two or more levels. A factorial design is currently not support for DESeq2.

For EDA, the input data does not need to have sample groups, and the input file can contain as few as two data columns. To identify DEGs, users should upload a file with at least two groups (biological samples) with replicates.

Pre-processing of counts data

Genes expressed at extremely low levels across all samples are excluded. By default, a gene must have at least 10 counts in one or more samples. This should be adjusted according to sequencing depth. There are 3 options for transformation of counts data for clustering analysis and PCA: variance stabilizing transform (VST), regularized log (rlog), and started log (slog). VST is performed according to [43] and rlog is part of DESeq2 [31]. When there are more than 10 samples, rlog becomes slow and is disabled. The default is slog, $log2(x+c)$, with a pseudo count c added to all counts before log transformation. The pseudo count can be changed within the range from 1 to 10. The bigger the pseudo count is, the less sensitive it is to noise from lowly expressed genes. For counts data, transformed data is only used for EDA, as the identification of

DEGs and pathway analysis are based on original counts. On the "Pre-processing" tab, users can download the processed data or search for the expression data for specific genes using gene symbols.

Pre-processing of FPKM or other normalized expression data

For normalized expression data, a filter is also applied to remove genes expressed at low levels in all samples. By default, only genes expressed at the level of 1 or higher in at least one sample are included in further analysis. This number should be adjusted according to the data distribution. For cDNA microarrays, where the expression levels are log ratios, we need to set this to a negative number such as -10^{20} to disable this filter.

iDEP calculates kurtosis for each of the data columns, and if the mean kurtosis is bigger than 50, a log2 transformation is enforced. Large kurtosis usually indicates the presence of extremely big numbers in the data set that warrants log-transformation. We believe such mechanisms make iDEP more robust, as users can forget to take log-transformation needed for highly skewed data, which can adversely affect all subsequent analyses. Users can double-check the effects of data transformation by examining the diagnostic plots generated on the Pre-processing tab.

Hierarchical clustering

Hierarchical clustering with heatmap can give us a holistic view of the data. iDEP first ranks all genes by standard deviation across all samples using the transformed data. By default, the top 1000 genes are used in hierarchical clustering with the average linkage method. The data is centered by subtracting the mean for each gene. The distance matrix is 1- *r*, where *r* is Pearson's correlation coefficient (PCC). Rendered by ggplot2 [59] and Plotly (https://plot.ly), the interactive heatmap enables users to zoom in to see gene symbols. A static heatmap with hierarchical clustering tree is also available. The correlation matrix is computed using the top 75% of genes ranked by average expression.

K-means clustering

k-Means clustering is an unsupervised method for clustering genes into groups based on their expression pattern across all samples. Genes are again ranked by standard deviation and a selected number of genes (2000 by default) are used for clustering. The default number of clusters is 4, which can be adjusted within 2 and 20. The data is normalized so that the rows (genes) have the same sum of absolute values (L1 norm).

Enrichment analysis is automatically conducted for each gene cluster using available gene sets. For each group, enriched TF binding motifs are identified. Transcript annotation and promoter sequences for 111 species are retrieved from Ensembl. For genes with multiple transcripts, the most common transcription start site (TSS) is used. If multiple TSS' have the same number of transcripts, then the most upstream TSS is used. Promoters are pre-scanned using TF binding motifs in CIS-BP [66]. Instead of defining a binary outcome of binding or not binding, which depends on arbitrary cutoffs, we recorded the best score for each of the TFs in every promoter sequences. Student's t-test is used to compare the scores observed in a group of genes against the rest. The P-values are corrected for multiple testing using false discovery rate (FDR).

Principal component analysis (PCA)

PCA is a linear transformation of data that enables us to project samples onto 2-dimensional (2D) spaces along the most variable directions. The plot can help visualize variations and grouping among samples. We also treated the PCA loadings onto each of the genes as expression data, and run pathway analysis with the PGSEA [35] package. This runs the PAGE algorithm [34], which performs one-sample t-tests on each gene set in the GO biological processes. The adjusted P-values are used to rank the pathways for each of the first 5 principal components. The pathways are labeled with FDR first, followed by the principal components (PC1, PC2 and so on). Only 5 pathways for each principal component are shown, but duplicated ones are

skipped. In addition to PCA, multidimensional scaling (MDS) is also available as a non-linear method to project samples onto 2D spaces while preserving the order of distances.

Differential gene expression

Linear Models for Microarray Data (*limma*) [30] is a widely used Bioconductor package for identifying DEGs. If the input is normalized expression data, *limma* is used to identify DEGs between all possible pairs of biological samples. A Venn diagram comparing all gene lists is produced. When there are more than 5 comparisons, only the first 5 gene lists are shown. In such cases, users can download the gene lists and produce Venn diagrams using other tools.

Users can examine each of comparisons through scatter and volcano plots, available as both static and interactive graphics. DEGs are used to conduct enrichment analysis using GO and other available gene sets. Similarly, enrichment of TF binding motif in promoters are conducted and the result is shown in a pop-up window. On the main panel, a heatmap is also produced showing the expression pattern and number of up- and down-regulated genes. The top genes among DEGs sorted by the absolute value of fold-changes are shown at the bottom of the main panel.

For read counts data, three methods are available to identify DEGs, namely DESeq2 [31], limma-voom [13,32], and limma-trend [13]. Even though slower than other methods, DESeq2 is set as default. Limma-trend and limma-voom first transform read counts data as continuous data. According to the *limma* user guide [67], limma-trend is robust when library sizes are similar; when there is more than 3-fold difference, limma-voom is recommended.

Pathway Analysis

Pathway analysis is done using fold-change values of all genes returned by *limma* or DESeq2. This choice disables sample permutations, but simplifies computation for different experiment designs. The sample sizes of many RNA-Seq studies are too small for sample permutation. GSEA (Gene Set Enrichment Analysis) [33] is conducted in the pre-ranked mode using a recently developed algorithm implemented in the *fgsea* package [68]. PAGE (Parametric Analysis of Gene Set Enrichment) [34] is used as implemented in *PGSEA* package [35]. Users can use PGSEA to analyze two selected biological samples in a comparison or across all samples in the study. The latter can produce a global view of significantly altered pathways across all conditions. GAGE (Generally Applicable Gene-set Enrichment) [36] is set as the default method for its speed and versatility. ReactomePA (Reactome Pathway Analysis) [37] is also included, which also uses the *fgsea* as the underlying algorithm. Unlike other methods, ReactomePA retrieves gene sets directly from Reactome pathway database [28,37], which focuses on human pathways but includes orthologous mappings for several model organisms.

The expression patterns of the genes belonging to significantly altered pathways can be visualized using heatmaps. For KEGG pathways [20], the Pathview [38] package is incorporated to show gene expression directly on pathway diagrams. Pathview downloads pathway diagrams directly from KEGG website and can be slow.

On the lower left side of the screen, there is check box named “Use absolute values of fold changes for GSEA and GAGE”. This is useful as some molecular pathways can be attenuated by up-regulating some of its genes while suppressing others. This is useful when using metabolic pathways such as KEGG. We should not check this when using TF or miRNA target genes, as the genes are often regulated collectively up or down.

A Plotly-based, interactive visualization of fold-change on all the chromosomes is shown on the “Chromosome” tab. Users can zoom in or mouse over to see gene symbols. iDEP also includes PREDA (Position RElated Data Analysis) package [39] that can detect chromosomal regions enriched with DEGs. PREDA uses nonlinear kernel regression to identify correlations between fold-changes and genomic

coordinates [69]. The 1000 permutations can require more than 5 minutes of computing, but it might be valuable for cancer-related expression studies when chromosomal abnormalities are possible.

Example RNA-Seq data on Hoxa1 knockdown in human fibroblasts is based on [3]. The raw data is available at Gene Expression Omnibus [70] with accession number GSE37704. Processed data is available as Supplementary File 3. Mouse Hoxa1 knockdown data is from [61] with accession number GSE50813 and available as Supplementary File 3. Expression data (GSE69210) of thymus tissues from patients with down syndrome are from [60] and the converted data is available as Supplementary File 4.

iDEP is available at http://ge-lab.org/idep. Online documentation, including frequently asked questions (FAQs), is available at https://idepsite.wordpress.com/. Source code is available at https://github.com/gexijin/iDEP. Gene set files for pathway analysis is available at http://ge-lab.org/#/data. Mailing list for user groups can be accessed at https://groups.google.com/d/forum/idep.

Acknowledgements

The author thanks Kevin Son for technical support of the web server. This work is partially supported by the National Science Foundation/EPSCoR Grant Number IIA-1355423 and by the State of South Dakota.

Conflicts of Interests

None.

References

1 Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods* **12**, 115-121, doi:10.1038/nmeth.3252 (2015).

2 Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, R36, doi:10.1186/gb-2013-14-4-r36 (2013).

3 Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* **31**, 46-53, doi:10.1038/nbt.2450 (2013).

4 Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**, 525-527, doi:10.1038/nbt.3519 (2016).

5 Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**, 417-419, doi:10.1038/nmeth.4197 (2017).

6 Reich, M. *et al.* GenePattern 2.0. *Nat Genet* **38**, 500-501, doi:10.1038/ng0506-500 (2006).

7 Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res* **44**, W3-W10, doi:10.1093/nar/gkw343 (2016).

8 Merchant, N. *et al.* The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences. *PLoS Biol* **14**, e1002342, doi:10.1371/journal.pbio.1002342 (2016).

9 Rstudio.

10 Nelson, J. W., Sklenar, J., Barnes, A. P. & Minnier, J. The START App: a web-based RNAseq analysis and visualization resource. *Bioinformatics*, doi:10.1093/bioinformatics/btw624 (2016).

11 Powell, D. R. <http://degust.erc.monash.edu/> (2016).

12 Dai, Z. *et al.* edgeR: a versatile tool for the analysis of shRNA-seq and CRISPR-Cas9 genetic screens. *F1000Res* **3**, 95, doi:10.12688/f1000research.3928.2 (2014).

13 Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **15**, R29, doi:10.1186/gb-2014-15-2-r29 (2014).

14 Pimentel, H., Bray, N., Puente, S., Melsted, P. & Pachter, L. in *BioRxiv* Vol. http://dx.doi.org/10.1101/058164 (2016).
15 Manning, J. *ShinyNGS*, <https://github.com/pinin4fjords/shinyngs> (2016).
16 Harshbarger, J., Kratz, A. & Carninci, P. DEIVA: a web application for interactive visual analysis of differential gene expression profiles. *BMC Genomics* **18**, 47, doi:10.1186/s12864-016-3396-5 (2017).
17 Younesy, H., Moller, T., Lorincz, M. C., Karimi, M. M. & Jones, S. J. VisRseq: R-based visual framework for analysis of sequencing data. *BMC Bioinformatics* **16 Suppl 11**, S2, doi:10.1186/1471-2105-16-S11-S2 (2015).
18 Gardeux, V., David, F., Shajkofci, A., Schwalie, P. & Deplancke, B. in *bioRxiv* (bioRxiv, http://biorxiv.org/content/early/2016/12/22/096222, 2016).
19 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29, doi:10.1038/75556 (2000).
20 Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* **45**, D353-D361, doi:10.1093/nar/gkw1092 (2017).
21 Zheng, H. Q. *et al.* EXPath tool-a system for comprehensively analyzing regulatory pathways and coexpression networks from high-throughput transcriptome data. *DNA Res*, doi:10.1093/dnares/dsx009 (2017).
22 Aken, B. L. *et al.* Ensembl 2017. *Nucleic Acids Res* **45**, D635-D642, doi:10.1093/nar/gkw1104 (2017).
23 Aken, B. L. *et al.* The Ensembl gene annotation system. *Database (Oxford)* **2016**, doi:10.1093/database/baw093 (2016).
24 Bolser, D. M., Staines, D. M., Perry, E. & Kersey, P. J. Ensembl Plants: Integrating Tools for Visualizing, Mining, and Analyzing Plant Genomic Data. *Methods Mol Biol* **1533**, 1-31, doi:10.1007/978-1-4939-6658-5_1 (2017).
25 Lai, e. a. GSKB: A gene set database for pathway analysis in mouse. *bioRxiv Preprint* http://biorxiv.org/content/early/2016/10/24/082511, 802511, doi:https://doi.org/10.1101/082511 (2016).
26 Lai, L. *et al.* AraPath: a knowledgebase for pathway analysis in Arabidopsis. *Bioinformatics* **28**, 2291-2292, doi:10.1093/bioinformatics/bts421 (2012).
27 Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417-425, doi:10.1016/j.cels.2015.12.004 (2015).
28 Fabregat, A. *et al.* The Reactome pathway Knowledgebase. *Nucleic Acids Res* **44**, D481-487, doi:10.1093/nar/gkv1351 (2016).
29 Ge, S. X. Exploratory bioinformatics investigation reveals importance of "junk" DNA in early embryo development. *BMC Genomics* **18**, 200, doi:10.1186/s12864-017-3566-0 (2017).
30 Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47, doi:10.1093/nar/gkv007 (2015).
31 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).
32 Liu, R. *et al.* Why weight? Modelling sample and observational level variability improves power in RNA-seq analyses. *Nucleic Acids Res* **43**, e97, doi:10.1093/nar/gkv412 (2015).
33 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550, doi:10.1073/pnas.0506580102 (2005).
34 Kim, S. Y. & Volsky, D. J. PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics* **6**, 144, doi:10.1186/1471-2105-6-144 (2005).

35 Furge, K. & Dykema, K. PGSEA: Parametric Gene Set Enrichment Analysis. *R package version 1.48.0.* (2012).
36 Luo, W., Friedman, M. S., Shedden, K., Hankenson, K. D. & Woolf, P. J. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics* **10**, 161, doi:10.1186/1471-2105-10-161 (2009).
37 Yu, G. & He, Q. Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol Biosyst* **12**, 477-479, doi:10.1039/c5mb00663e (2016).
38 Luo, W. & Brouwer, C. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics* **29**, 1830-1831, doi:10.1093/bioinformatics/btt285 (2013).
39 Ferrari, F., Solari, A., Battaglia, C. & Bicciato, S. PREDA: an R-package to identify regional variations in genomic data. *Bioinformatics* **27**, 2446-2447, doi:10.1093/bioinformatics/btr404 (2011).
40 Turner, S. (http://www.gettinggeneticsdone.com/2015/12/tutorial-rna-seq-differential.html, 2015).
41 Patro, R., Mount, S. M. & Kingsford, C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat Biotechnol* **32**, 462-464, doi:10.1038/nbt.2862 (2014).
42 Himes, B. E. *et al.* RNA-Seq transcriptome profiling identifies CRISPLD2 as a glucocorticoid responsive gene that modulates cytokine function in airway smooth muscle cells. *PLoS One* **9**, e99625, doi:10.1371/journal.pone.0099625 (2014).
43 Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol* **11**, R106, doi:10.1186/gb-2010-11-10-r106 (2010).
44 Nahle, Z. *et al.* Direct coupling of the cell cycle and cell death machinery by E2F. *Nat Cell Biol* **4**, 859-864, doi:10.1038/ncb868 (2002).
45 Liang, H. *et al.* Neural development is dependent on the function of specificity protein 2 in cell cycle progression. *Development* **140**, 552-561, doi:10.1242/dev.085621 (2013).
46 DeGregori, J. The genetics of the E2F family of transcription factors: shared functions and unique roles. *Biochim Biophys Acta* **1602**, 131-150 (2002).
47 Xie, X. *et al.* Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* **434**, 338-345, doi:10.1038/nature03441 (2005).
48 Cloonan, N. *et al.* The miR-17-5p microRNA is a key regulator of the G1/S phase cell cycle transition. *Genome Biol* **9**, R127, doi:10.1186/gb-2008-9-8-r127 (2008).
49 Trompeter, H. I. *et al.* MicroRNAs MiR-17, MiR-20a, and MiR-106b act in concert to modulate E2F activity on cell cycle arrest during neuronal lineage differentiation of USSC. *PLoS One* **6**, e16138, doi:10.1371/journal.pone.0016138 (2011).
50 Petrocca, F. *et al.* E2F1-regulated microRNAs impair TGFbeta-dependent cell-cycle arrest and apoptosis in gastric cancer. *Cancer Cell* **13**, 272-286, doi:10.1016/j.ccr.2008.02.013 (2008).
51 Paraguison, R. C. *et al.* Enhanced autophagic cell death in expanded polyhistidine variants of HOXA1 reduces PBX1-coupled transcriptional activity and inhibits neuronal differentiation. *J Neurosci Res* **85**, 479-487, doi:10.1002/jnr.21137 (2007).
52 Gavalas, A., Ruhrberg, C., Livet, J., Henderson, C. E. & Krumlauf, R. Neuronal defects in the hindbrain of Hoxa1, Hoxb1 and Hoxb2 mutants reflect regulatory interactions among these Hox genes. *Development* **130**, 5663-5679, doi:10.1242/dev.00802 (2003).
53 Canu, E. *et al.* HOXA1 A218G polymorphism is associated with smaller cerebellar volume in healthy humans. *J Neuroimaging* **19**, 353-358, doi:10.1111/j.1552-6569.2008.00326.x (2009).
54 Chang, H. Y. *et al.* Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. *PLoS Biol* **2**, E7, doi:10.1371/journal.pbio.0020007 (2004).

55 Ge, S. X. Large-scale analysis of expression signatures reveals hidden links among diverse cellular processes. *BMC Syst Biol* **5**, 87, doi:10.1186/1752-0509-5-87 (2011).
56 Vermeulen, K., Van Bockstaele, D. R. & Berneman, Z. N. The cell cycle: a review of regulation, deregulation and therapeutic targets in cancer. *Cell Prolif* **36**, 131-149 (2003).
57 Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307-319, doi:10.1016/j.cell.2013.03.035 (2013).
58 Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380, doi:10.1038/nature11082 (2012).
59 Wickham, H. *Ggplot2 : elegant graphics for data analysis*. (Springer, 2009).
60 Moreira-Filho, C. A. *et al.* Modular transcriptional repertoire and MicroRNA target analyses characterize genomic dysregulation in the thymus of Down syndrome infants. *Oncotarget* **7**, 7497-7533, doi:10.18632/oncotarget.7120 (2016).
61 Brock, A. *et al.* Silencing HoxA1 by intraductal injection of siRNA lipidoid nanoparticles prevents mammary tumor progression in mice. *Sci Transl Med* **6**, 217ra212, doi:10.1126/scitranslmed.3007048 (2014).
62 Ioannidis, J. P. Why most published research findings are false. *PLoS Med* **2**, e124, doi:10.1371/journal.pmed.0020124 (2005).
63 Smedley, D. *et al.* The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res* **43**, W589-598, doi:10.1093/nar/gkv350 (2015).
64 Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439-3440, doi:10.1093/bioinformatics/bti525 (2005).
65 Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc* **4**, 1184-1191, doi:10.1038/nprot.2009.97 (2009).
66 Weirauch, M. T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431-1443, doi:10.1016/j.cell.2014.08.009 (2014).
67 Smyth, G. K. *limma: Linear Models for Microarray and RNA-Seq Data: User's Guide*, <https://bioconductor.org/packages/release/bioc/vignettes/limma/inst/doc/usersguide.pdf> (2016).
68 Sergushichev, A. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *bioRxiv* http://biorxiv.org/content/early/2016/06/20/060012, doi:doi: 10.1101/060012 (2016).
69 Callegaro, A., Basso, D. & Bicciato, S. A locally adaptive statistical procedure (LAP) to identify differentially expressed chromosomal regions. *Bioinformatics* **22**, 2658-2666, doi:10.1093/bioinformatics/btl455 (2006).
70 Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**, 207-210 (2002).

Tables

Table 1. Enriched Gene Ontology terms for clusters of genes defined by k-means clustering.

Cluster	adj.Pval	Genes	Pathways
A	4.3E-54	252	Cell cycle
	1.4E-53	219	Cell cycle process
	4.2E-53	182	Mitotic cell cycle
	1.7E-52	174	Mitotic cell cycle process
	1.1E-39	119	Nuclear division
	2.5E-39	103	Mitotic nuclear division
	4.2E-39	90	Chromosome segregation
	5.6E-39	126	Chromosome organization
	4.8E-38	120	Organelle fission
	1.9E-36	71	Sister chromatid segregation
B	2.8E-18	143	Response to endogenous stimulus
	2.8E-18	125	Cell migration
	2.8E-18	87	Regulation of cell migration
	2.8E-18	144	Locomotion
	2.8E-18	94	Regulation of locomotion
	3.0E-18	145	Regulation of cell proliferation
	3.5E-18	90	Regulation of cell motility
	3.9E-18	94	Regulation of cellular component movement
	4.3E-18	131	Cell motility
	4.3E-18	131	Localization of cell

Table 2. Enriched TF binding motifs in the promoters (300bp upstream) of gene clusters defined by k-means clustering.

List	Motif	TF	TF family	FDR
A	CACGTG	TCFL5	bHLH	0.0E+00
	GGCGGGAA	E2F4	E2F	0.0E+00
	GGGGGCGGGGC	SP2	C2H2 ZF	0.0E+00
	GGCCGGAG	MBD2	MBD	0.0E+00
	TGCGGG	ZBTB1	C2H2 ZF	2.8E-14
	GTGGGCGTGGC	SP6	C2H2 ZF	4.9E-13
	GGGCGTG	KLF7	C2H2 ZF	8.7E-13
	GGGCGGGAA	E2F6	E2F	5.7E-11
	ATGCGTGGGCGG	EGR4	C2H2 ZF	1.0E-10
	GCGCCAAA	E2F5	E2F	2.3E-10
B	CACGTG	TCFL5	bHLH	2.6E-10
	TGCGGG	ZBTB1	C2H2 ZF	2.6E-10
	GGGGGCGGGGC	SP2	C2H2 ZF	3.1E-09
	GGGGGGGGGCC	PATZ1	C2H2 ZF	5.5E-09
	GGCCGGAG	MBD2	MBD	1.5E-07
	CACAGCGGGGGGTC	ZIC4	C2H2 ZF	2.2E-06
	GGGCGGGAA	E2F6	E2F	1.5E-05
	GGGGGGT	ZIC5	C2H2 ZF	1.5E-05
	GGCGGGAA	E2F4	E2F	1.6E-05
	GTGGGCGTGG	SP8	C2H2 ZF	2.0E-05

Table 3. Enriched GO Biological Process terms in differential expressed genes.

Direction	adj.Pval	Genes	Pathways
Hoxa1KN > control	2.5E-24	161	Regulation of cell proliferation
	1.7E-18	145	Response to endogenous stimulus
	1.7E-18	146	Locomotion
	5.4E-18	93	Regulation of locomotion
	1.2E-17	94	Regulation of cellular component movement
Hoxa1KN < control	2.6E-47	239	Cell cycle
	2.6E-47	208	Cell cycle process
	2.6E-47	173	Mitotic cell cycle
	1.4E-46	165	Mitotic cell cycle process
	7.7E-38	101	Mitotic nuclear division

Table 4. Enriched TF binding motifs in the promoters (300bp upstream) of differential expressed genes.

List	Motif	TF	TF family	FDR
Hoxa1KN > control	GGGGGGGGGCC	PATZ1	C2H2 ZF	2.4E-5
	GGGGGCGGGGC	SP2	C2H2 ZF	2.4E-5
	TGCGGG	ZBTB1	C2H2 ZF	2.3E-4
	GGCCGGAG	MBD2	MBD	3.5E-4
	GGGGGGT	ZIC5	C2H2 ZF	6.6E-4
	CACGTG	TCFL5	bHLH	1.9E-3
	CACAGCGGGGGGTC	ZIC4	C2H2 ZF	4.1E-3
Hoxa1KN < control	GGCGGGAA	E2F4	E2F	2.8E-12
	GGCCGGAG	MBD2	MBD	2.8E-12
	CACGTG	TCFL5	bHLH	1.1E-10
	GGGGGCGGGGC	SP2	C2H2 ZF	1.4E-9
	GGGCGGGAA	E2F6	E2F	2.4E-8
	GTGGGCGTGGC	SP6	C2H2 ZF	4.1E-8
	GCGCCAAA	E2F5	E2F	5.5E-8
	TGCGGG	ZBTB1	C2H2 ZF	1.3E-7
	GGGCGTG	KLF7	C2H2 ZF	2.2E-7
	ATGCGTGGGCGG	EGR4	C2H2 ZF	3.8E-6
	CACAGCGGGGGGTC	ZIC4	C2H2 ZF	4.6E-6
	GTGGGCGTGG	SP8	C2H2 ZF	1.4E-5
	TTTGCGCC	MYPOP	Myb/SANT	1.5E-5
	AAATGGCGGGAAA	TFDP2	DP,E2F	9.5E-5
	CCCGCATACAACGAA	CENPB	CENPB	2.4E-4
	GGGGGGGGGCC	PATZ1	C2H2 ZF	2.5E-4
	GCCAATCA	PBX3	Homeodomain	3.8E-4

Table 5. Overrepresented MSigDB motif gene sets in differential expressed genes.

Direction	adj.Pval	Genes	Pathways
Hoxa1KN > control	4.6E-14	152	AACTTT UNKNOWN
	1.4E-13	159	TTGTTT V$FOXO4 01
	1.2E-11	175	CAGGTG V$E12 Q6
	1.2E-11	110	TATAAA V$TATA 01
	1.1E-09	137	TGGAAA V$NFAT Q4 01
	7.0E-09	138	CTTTGT V$LEF1 Q2
	8.7E-09	80	RTAAACA V$FREAC2 01
	6.0E-08	58	GCACTTT MIR-17-5P, MIR-20A, MIR-106A, MIR-106B, MIR-20B, MIR-519D
	1.4E-07	36	TATTATA,MIR-374
	1.8E-06	144	GGGAGGRR V$MAZ Q6
Hoxa1KN < control	2.0E-09	39	V$E2F 02
	2.0E-09	39	V$E2F Q4
	2.0E-09	39	V$E2F Q6
	2.0E-09	40	V$E2F1 Q6
	2.0E-09	39	V$E2F1DP1 01
	2.0E-09	39	V$E2F1DP2 01
	2.0E-09	39	V$E2F4DP2 01
	2.0E-09	39	V$E2F1DP1RB 01
	2.4E-09	39	V$E2F4DP1 01
	2.4E-09	32	SGCGSSAAA V$E2F1DP2 01

Table 6. Enriched GO Cellular Component terms in the differentially expressed genes.

Direction	adj.Pval	Genes	Pathways
control < Hoxa1KN	3.7E-12	143	Intrinsic component of plasma membrane
	6.1E-11	135	Integral component of plasma membrane
	7.2E-07	104	Extracellular space
	3.0E-06	75	Plasma membrane region
	7.3E-06	93	Neuron part
	6.1E-05	36	Proteinaceous extracellular matrix
	3.5E-04	45	Extracellular matrix
	3.5E-04	67	Neuron projection
	3.5E-04	57	Synapse
	3.9E-04	28	Synaptic membrane
control > Hoxa1KN	1.0E-25	70	Spindle
	2.9E-24	130	Chromosome
	2.9E-24	144	Microtubule cytoskeleton
	4.3E-24	54	Chromosome, centromeric region
	4.4E-24	56	Condensed chromosome
	6.8E-24	178	Cytoskeletal part
	3.1E-23	73	Chromosomal region
	1.8E-22	40	Condensed chromosome, centromeric region
	2.2E-22	43	Kinetochore
	1.6E-20	112	Chromosomal part

Table 7. Pathways detected by applying GAGE method to KEGG pathways.

Direction	GAGE analysis: Hoxa1KN vs control	statistic	Genes	adj.Pval
Hoxa1KN < control	Cell cycle	-4.7765	113	0.00045
	DNA replication	-4.6424	32	0.0012
	Oocyte meiosis	-3.7556	100	0.01
	Homologous recombination	-3.2536	35	0.042
	Pyrimidine metabolism	-3.2051	93	0.042
	Biosynthesis of amino acids	-3.1759	56	0.042
Hoxa1KN > control	Cytokine-cytokine receptor interaction	4.5247	127	0.0014
	Rheumatoid arthritis	3.5479	49	0.043
	Lysosome	3.3444	107	0.043
	Hematopoietic cell lineage	2.8896	39	0.17

Table 8. Undirected pathway analysis using the absolute values of fold changes.

Direction	GAGE analysis: Hoxa1KN vs control	statistic	Genes	adj.Pval
Hoxa1KN < control	Ribosome	-5.3671	121	3.9E-05
	Huntington's disease	-4.9352	166	9.4E-05
	Parkinson's disease	-4.6964	124	2.2E-04
	Oxidative phosphorylation	-4.5404	115	3.1E-04
	Spliceosome	-4.5144	120	3.1E-04
	mRNA surveillance pathway	-3.3987	75	2.1E-02
	Proteasome	-3.3342	35	3.4E-02
	Protein processing in endoplasmic reticulum	-2.9453	145	5.7E-02
	Ribosome biogenesis in eukaryotes	-2.825	69	8.3E-02
Hoxa1KN > control	Cytokine-cytokine receptor interaction	4.6618	127	7.7E-04
	Cell cycle	3.5968	113	2.0E-02
	Rheumatoid arthritis	3.5965	49	2.0E-02
	Neuroactive ligand-receptor interaction	3.4877	105	2.0E-02
	Pathways in cancer	3.3261	323	2.4E-02
	Cell adhesion molecules (CAMs)	3.0643	74	5.7E-02
	NF-kappa B signaling pathway	2.9582	68	6.4E-02
	HTLV-I infection	2.9036	196	6.4E-02
	DNA replication	2.8474	32	8.6E-02

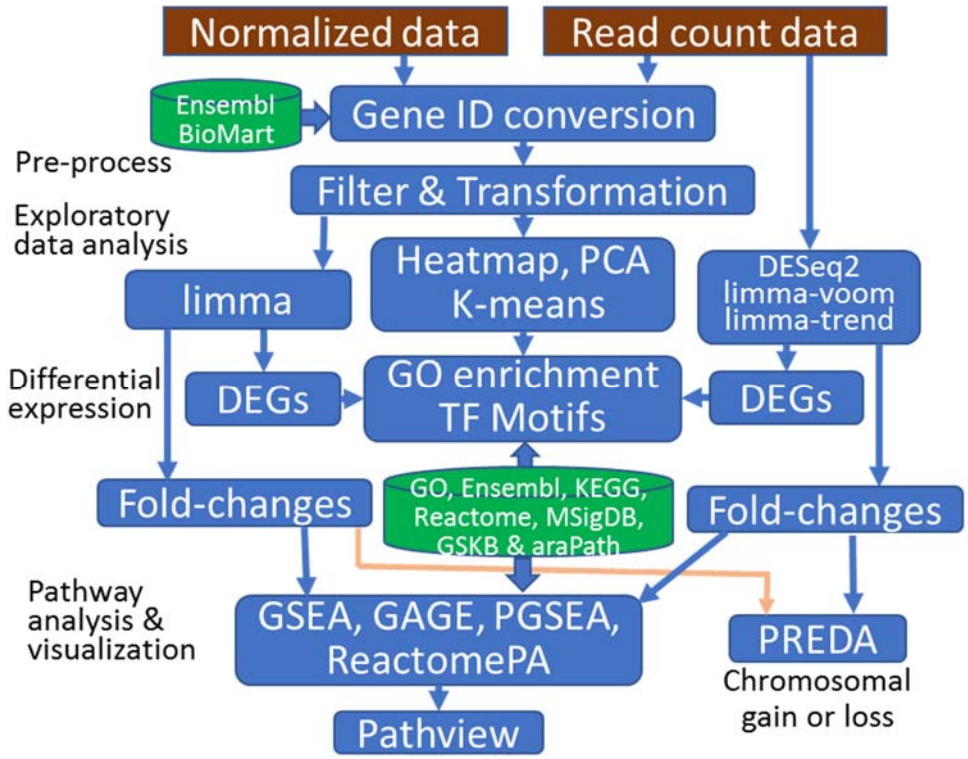


Figure 1. Flowchart for iDEP. Integration of various R and Bioconductor packages with annotation databases enables analyses of gene expression data through graphical user interface.

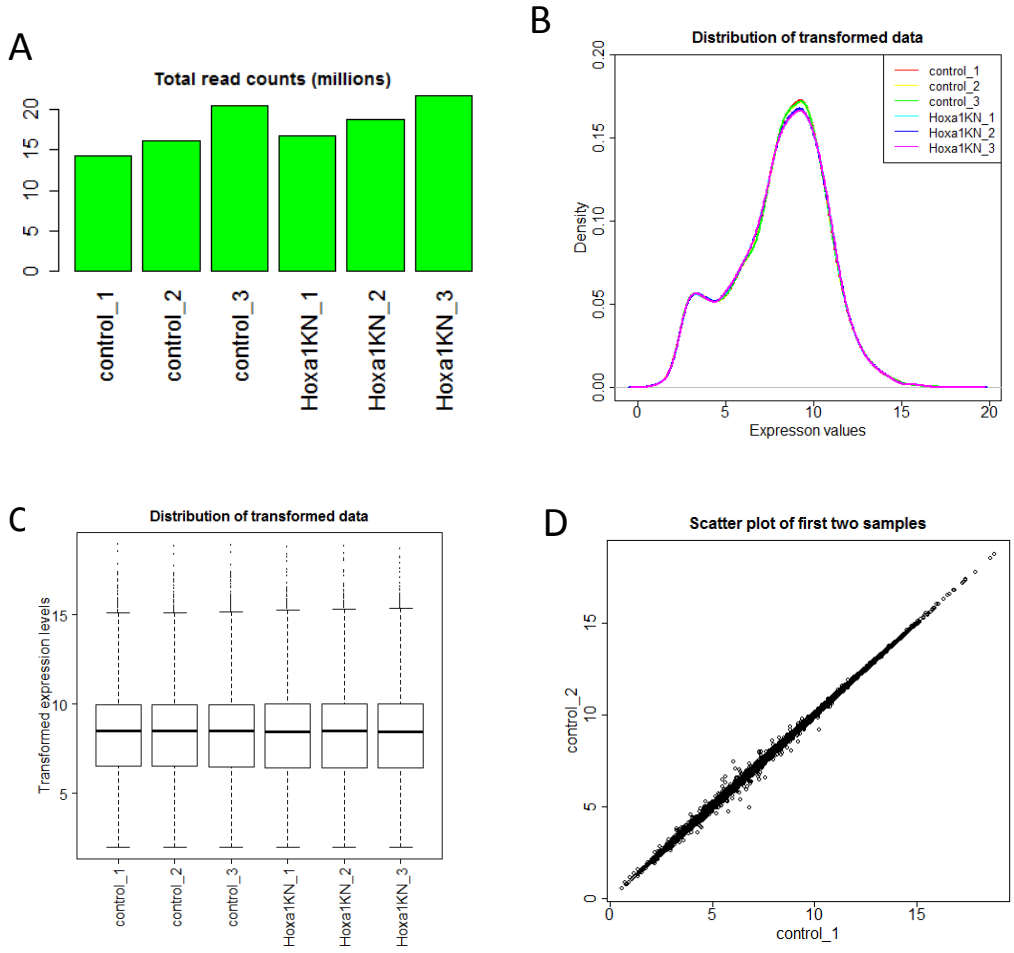


Figure 2. Exploratory data analysis for RNA-Seq read counts data of Hoxa1 knockdown. A) Total reads per library. B) Density plots of transformed read-counts data. C) Box plots of transformed data. D) Scatter plot of technical replicates.

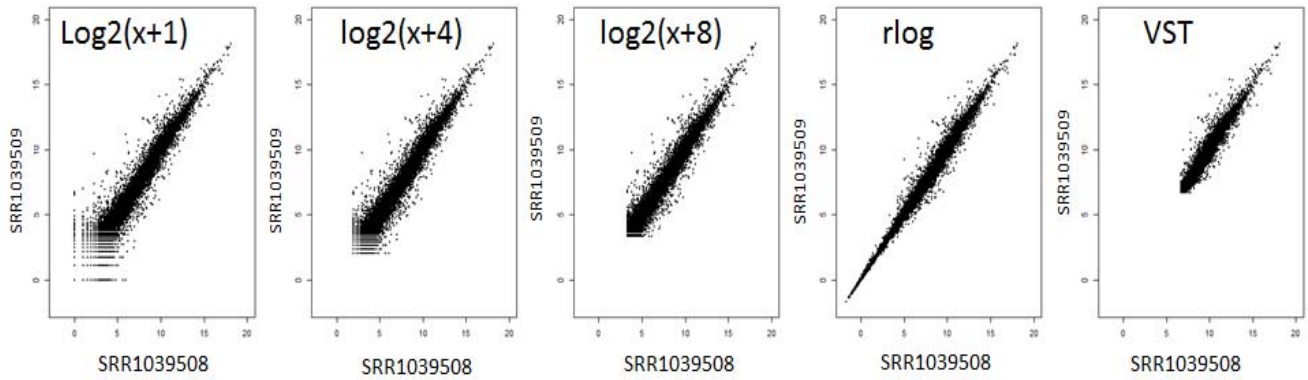


Figure 3. Effect of various data transformations on RNA-Seq read count data. The two libraries are from different biological samples. Transformations drastically alter lowly expressed genes.

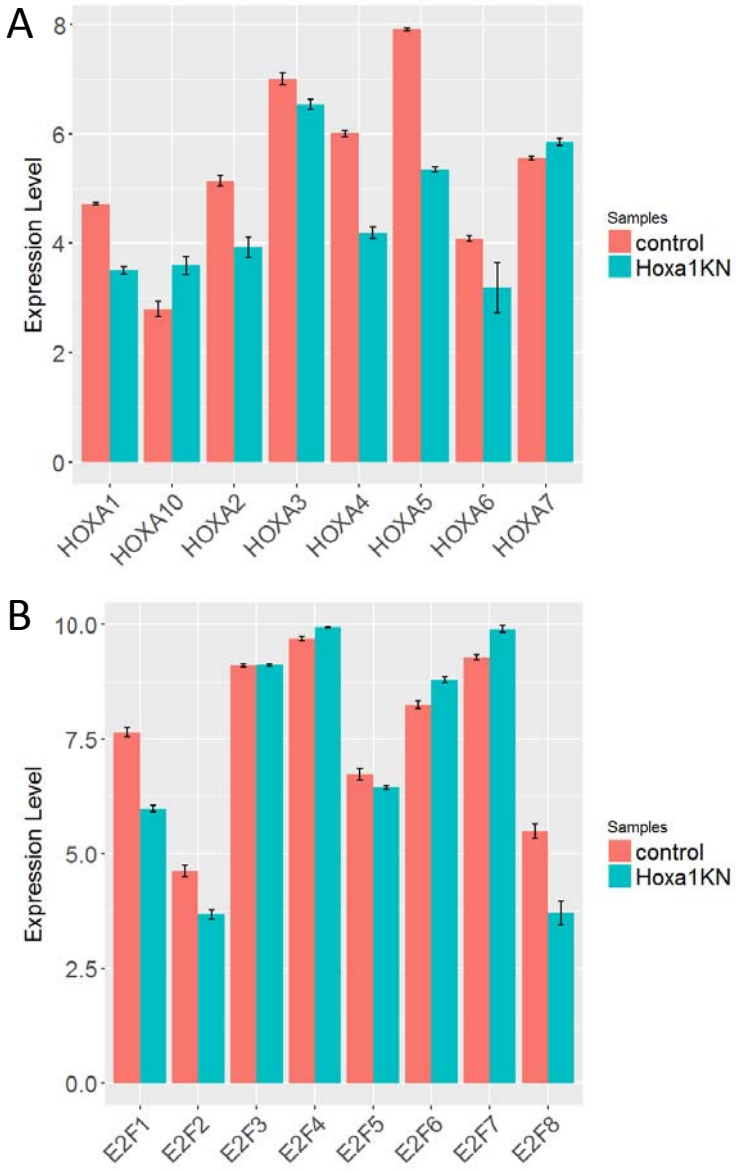


Figure 4. Expression levels of Hoxa (A) and E2F (B) family genes, generated by iDEP using keyword searches.

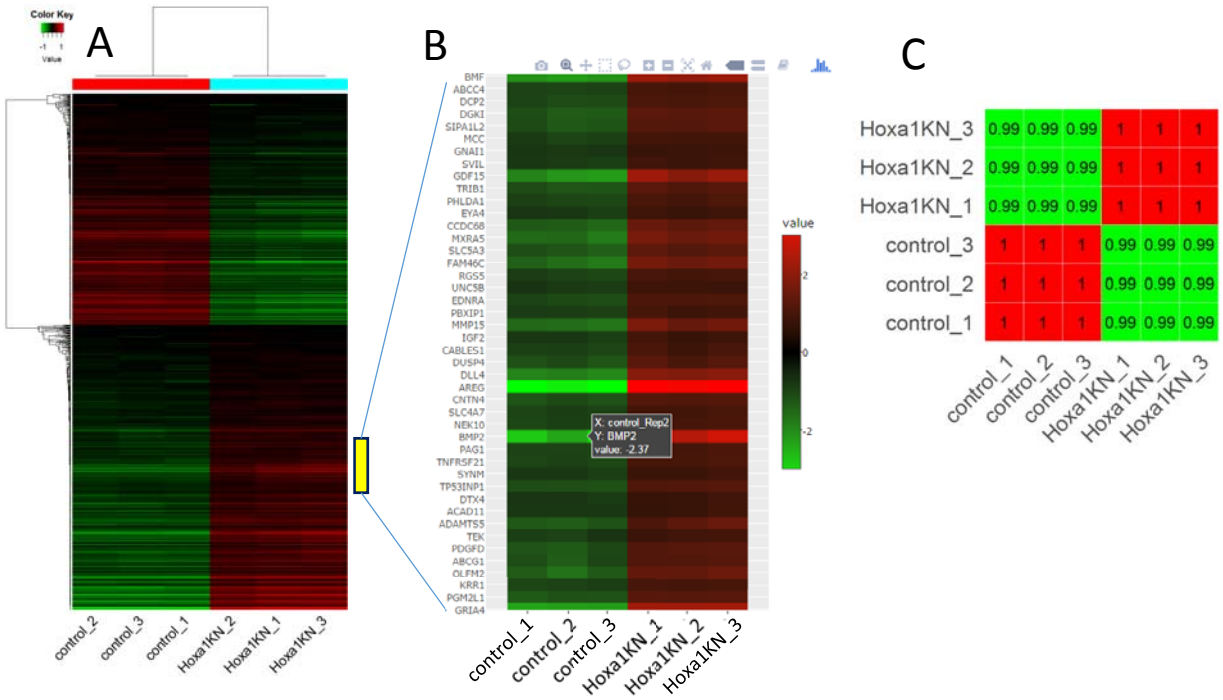


Figure 5. Results of hierarchical clustering (A), zoom-in using the interactive heatmap (B) and correlation matrix (C).

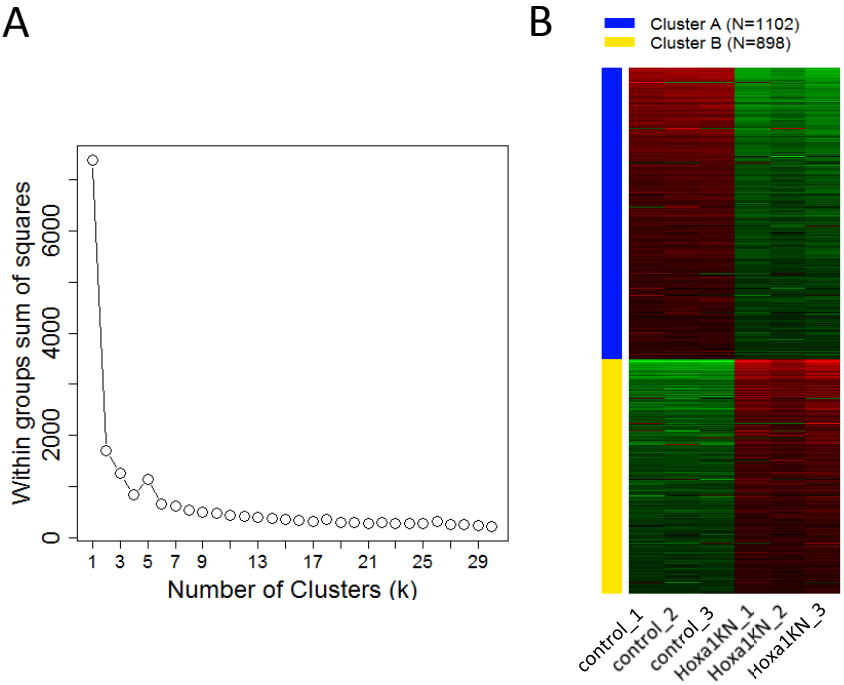


Figure 6. Results from K-means clustering. A) Plotting of the within groups sum of squares to guide the choices of number of clusters. B) Using k=2, iDEP classified top 2000 genes into 2 groups.

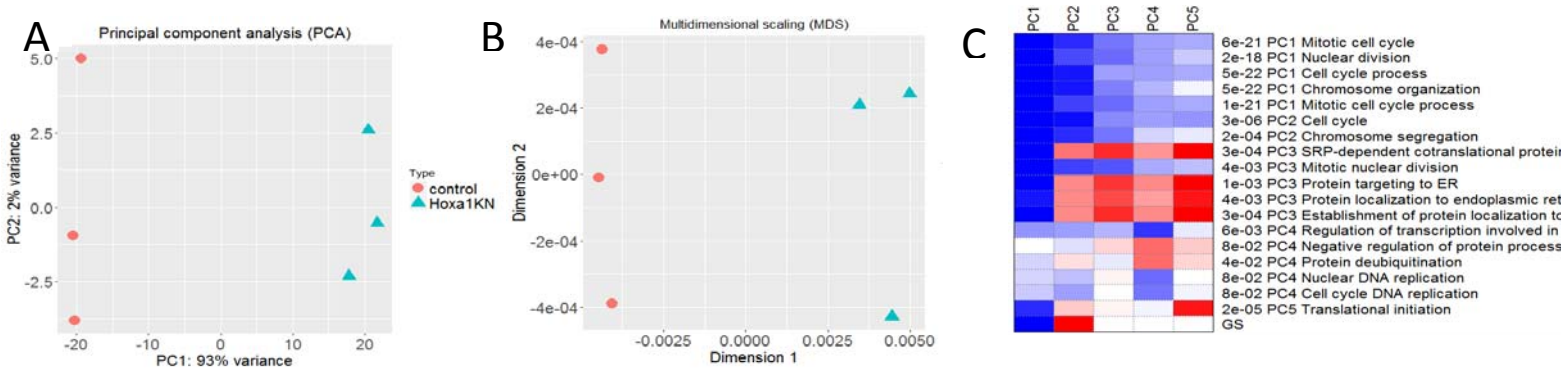


Figure 7. Results from PCA (A) and MDS (B). C. Pathway analysis using PGSEA on the PCA loadings.

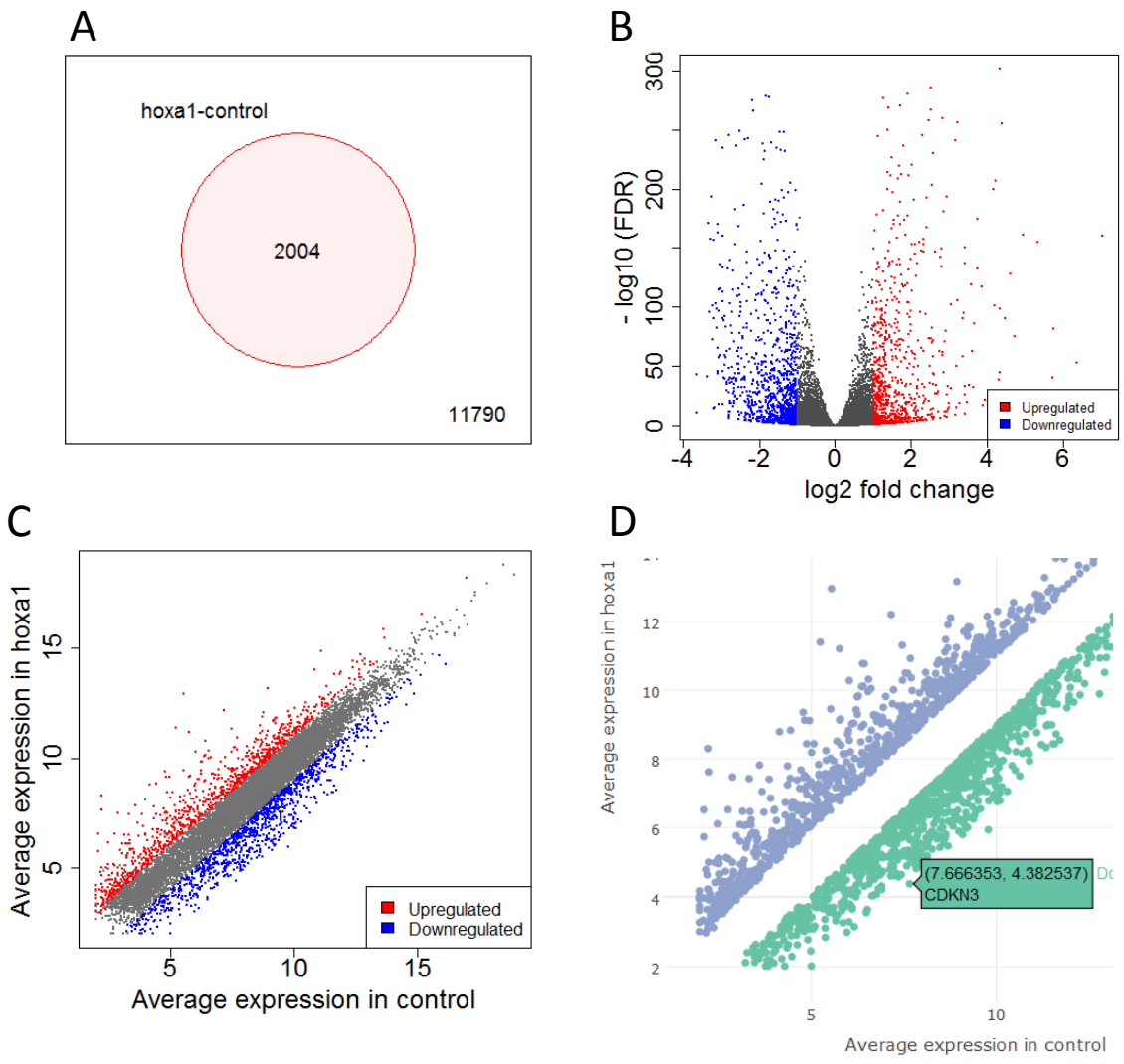


Figure 8. Summary plots for differential expression analysis using DESeq2. A) Venn diagram. Out of the 11,790 genes, 2004 are differentially expressed. B) Volcano plot, C) Scatter plot, and D) zoom-in of the interactive scatter plot.

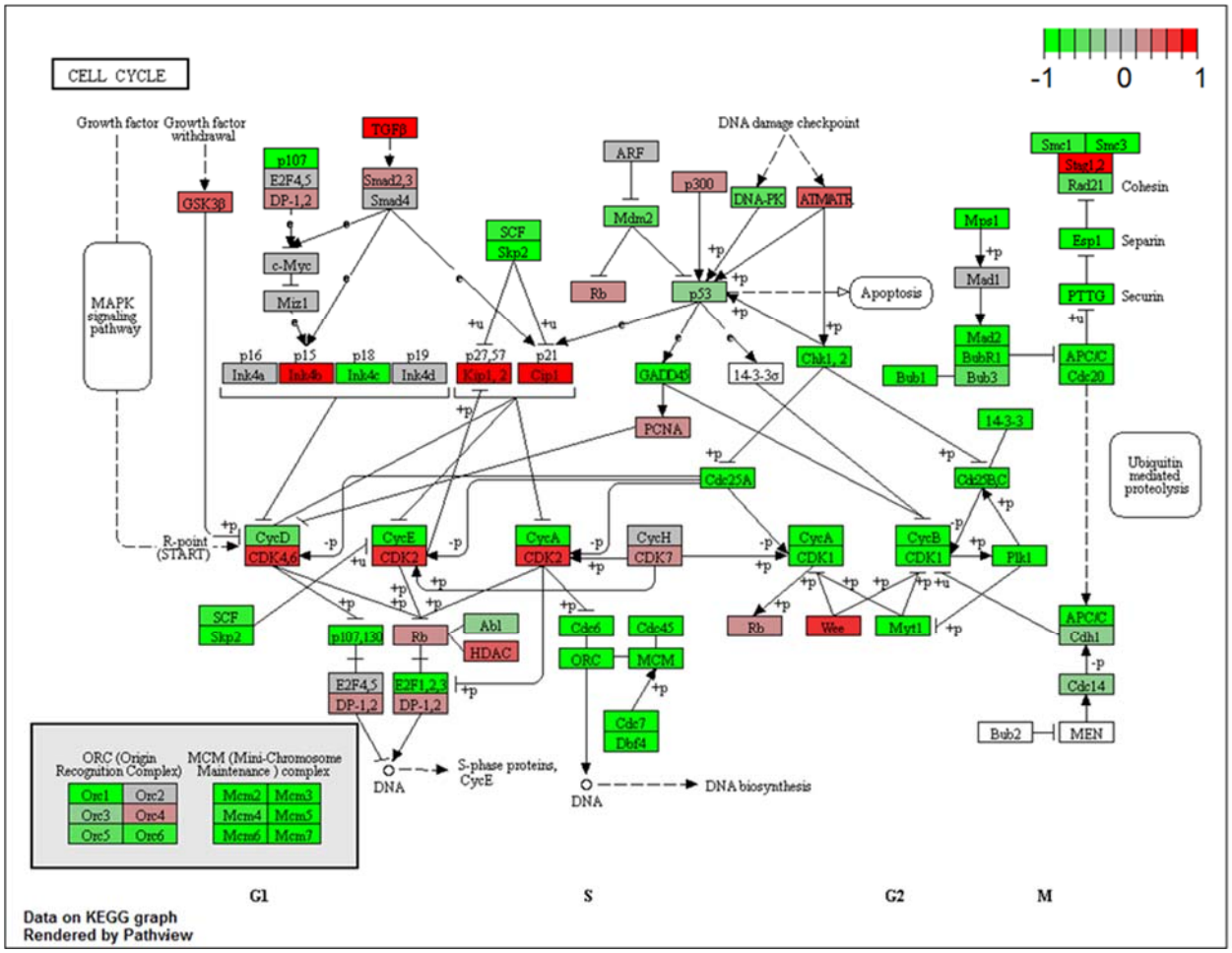


Figure 9. Expression profiles of cell cycle related genes visualized on KEGG pathway diagram. This is created using the Pathview package included in iDEP. Red and green indicate genes induced or suppressed by Hoxa1 knockdown, respectively.

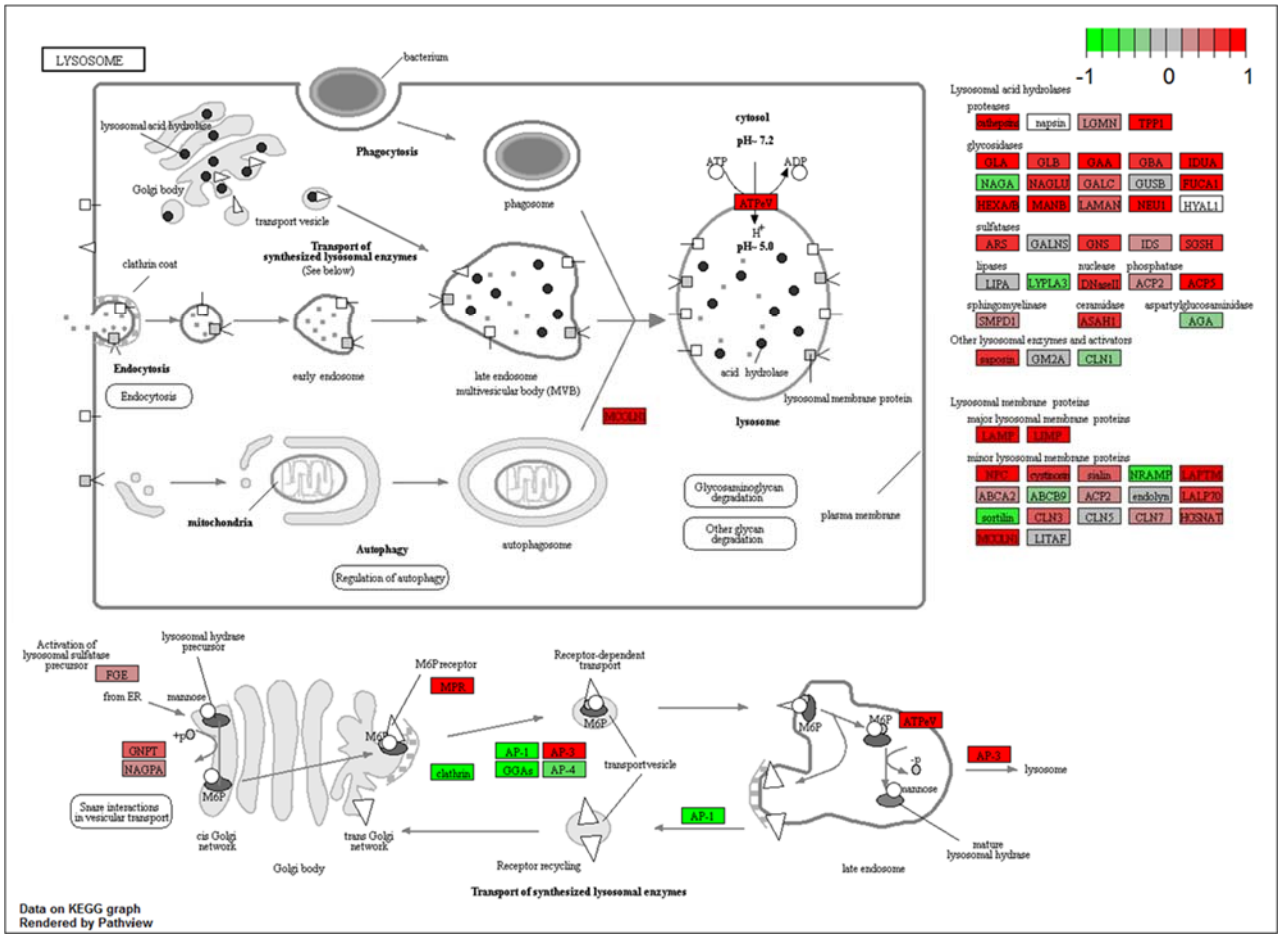


Figure 10. Gene related to lysosome are upregulated (red) in Hoxa1 knockdown samples.

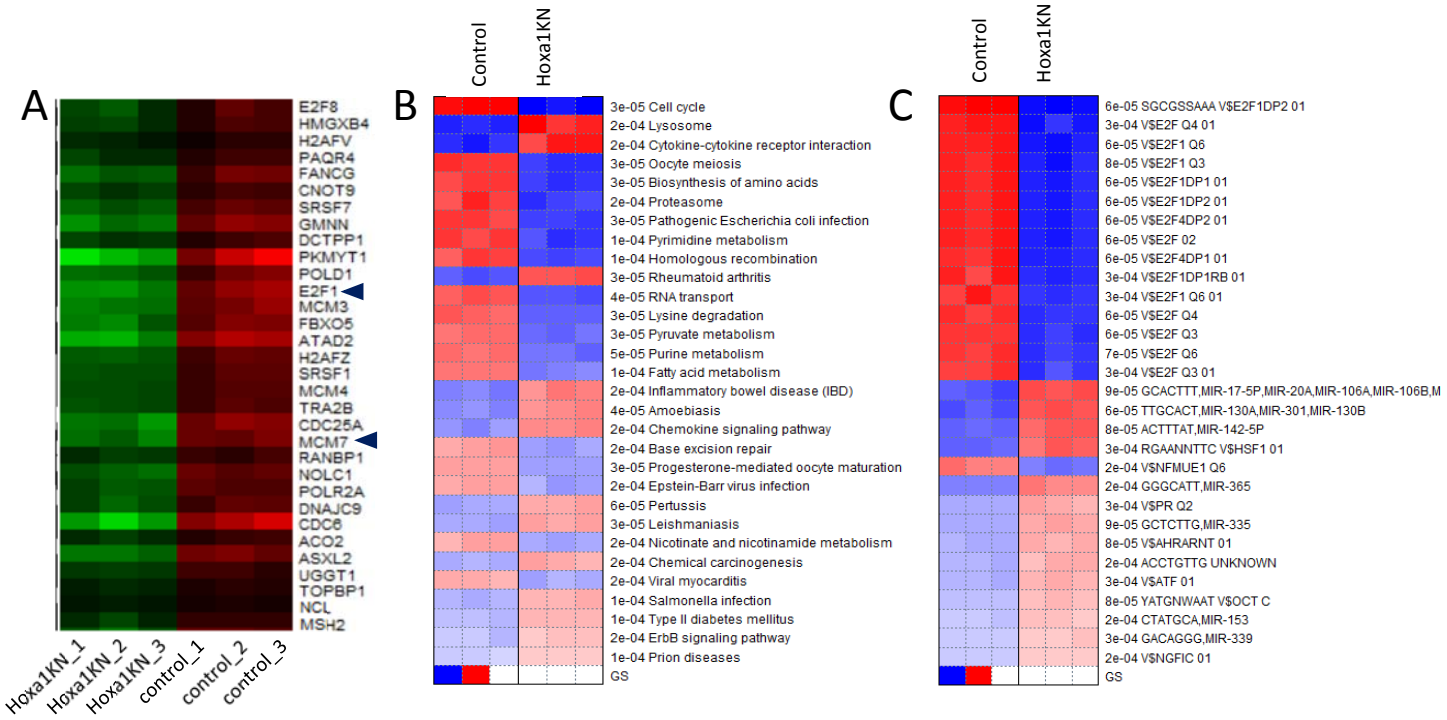


Figure 11. Pathway analysis results using different options. (A) expression patterns of genes with E2F1 binding motifs. E2F1 gene itself is also downregulated in Hoxa1 knockdown. So is the Mcm7 gene, whose intron host miR-106b-25 clusters. B) Results from running PGSEA on KEGG gene sets. C) PGSEA applied on MSigDB.Motif gene sets.

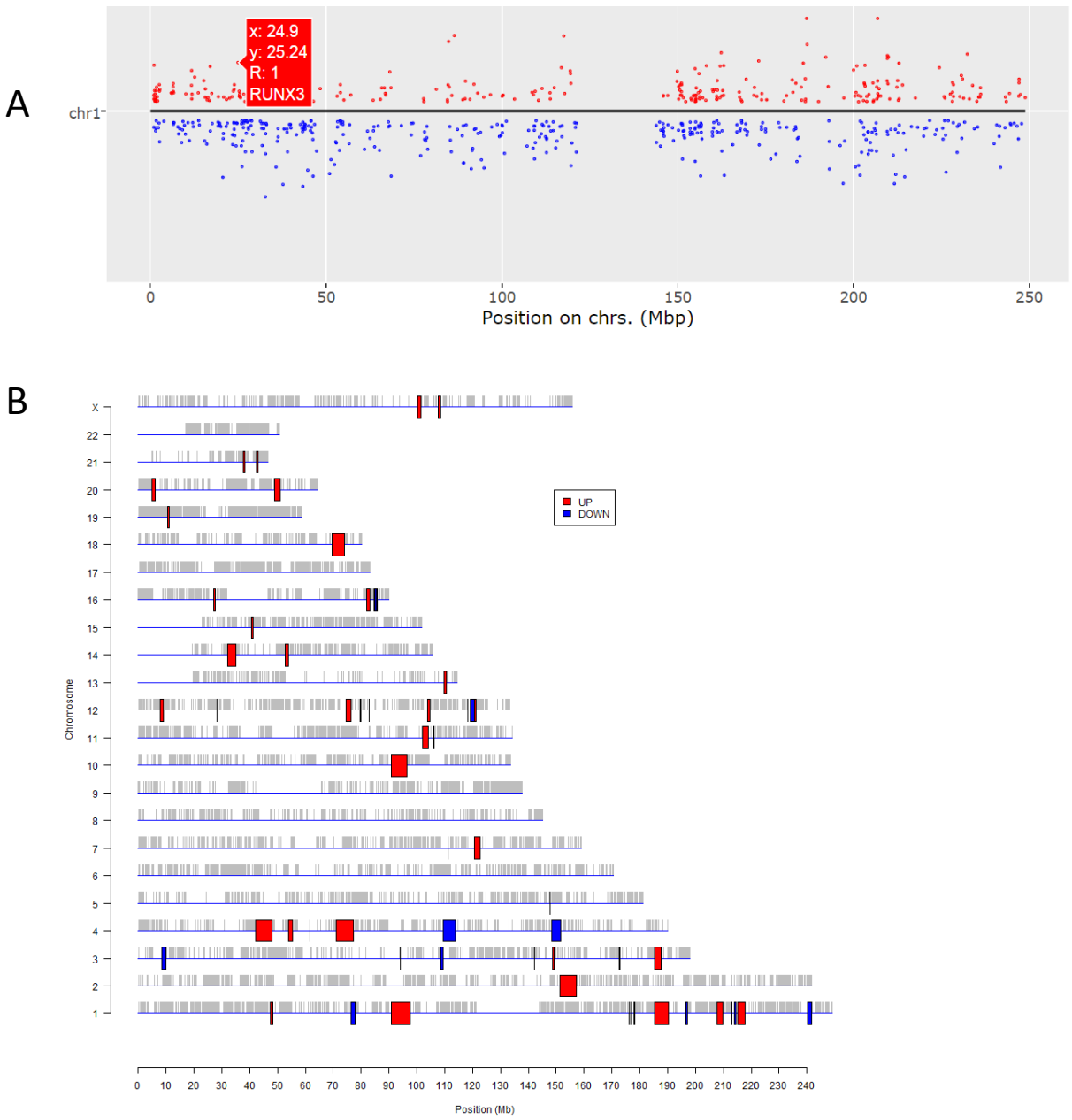


Figure 12. Visualizing expression profiles on chromosomes. A) Zoom-in on Chr. 1 using the dynamic graphics, showing the upregulation of RUX3 gene. B) Statistically significant genomic regions identified by PREDA (FDR<.01).

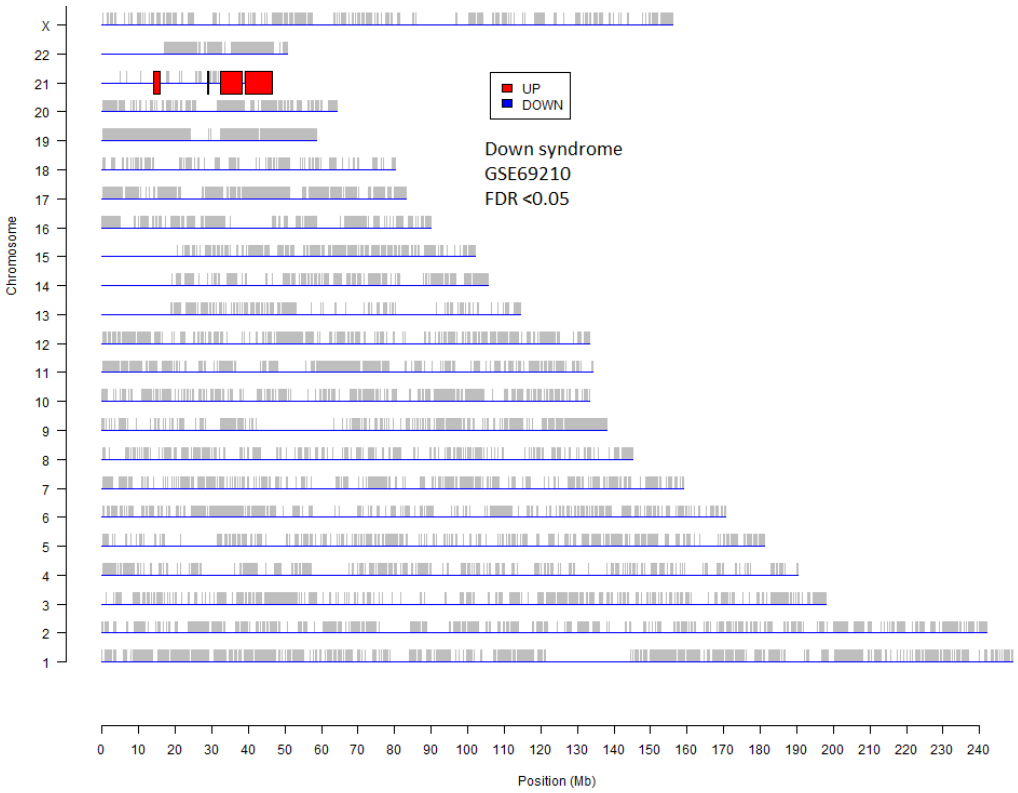
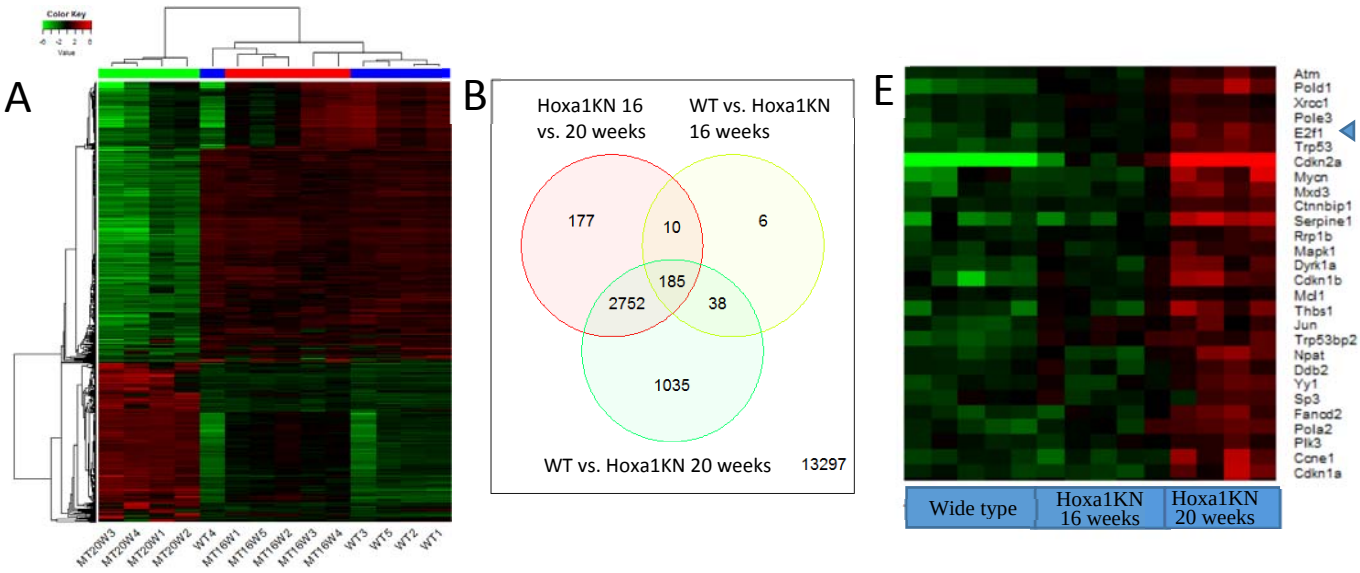


Figure 13. PREDA detects upregulated chromosomal regions on Chr. 21 in samples with down syndrome.



C

Direction	GSEA analysis: WT vs MT20W	NES	Genes	adj.Pval
WT < MT20W	Cell cycle	-2.6034	120	3.0e-03
	Spliceosome	-2.5338	118	3.0e-03
	DNA replication	-2.5315	32	3.0e-03
	Homologous recombination	-2.2348	37	3.0e-03
	Mismatch repair	-2.2018	18	3.0e-03
	Fanconi anemia pathway	-2.195	45	3.0e-03
	RNA transport	-2.1776	152	3.0e-03
	Nucleotide excision repair	-2.0994	40	3.0e-03
	TNF signaling pathway	-2.0425	103	3.0e-03
	P53 signaling pathway	-2.0366	65	3.0e-03
	MRNA surveillance pathway	-1.9656	83	3.0e-03
WT > MT20W	PPAR signaling pathway	2.5839	75	3.0e-03
	Valine, leucine and isoleucine degradation	2.556	50	3.0e-03

D

Direction	GAGE analysis: WT vs MT20W	statistic	Genes	adj.Pval
WT < MT20W	TFACTS binding motif:E2F1	-5.311	145	2.4e-05
	TFACTS binding motif:E2F3	-4.9515	34	3.9e-04
	TFACTS binding motif:E2F2	-4.881	31	4.0e-04
	TFACTS binding motif:E2F4	-4.3078	35	1.6e-03
	TFACTS binding motif:ATF1	-2.8962	54	8.5e-02
WT > MT20W	TFACTS binding motif:SREBF1	3.556	47	6.3e-02
	TFACTS binding motif:PPARA	3.2254	19	9.8e-02
	TF binding motif:ZHAO AR	3.0076	70	9.8e-02

Figure 14. Results from analyzing mouse gene expression study of Hoxa1 knockdown using siRNA. A) Hierarchical clustering indicates one outlier sample ("WT4"). B) Venn diagram showing the overlap of three lists of DEGs derived from three comparisons. C) Results of pathway analysis using GSEA. D) Enrichment of E2F motifs is also found. E) Expression of E2F1 and its target genes increased in Hoxa1 knockdown, contrary to what observed in human.

Email us for questions, suggestions, or data contributions. Stay connected via user group, or Twitter.

R as in Reproducibility

We recommend users to save the following details about their analyses.
Analysis were conducted using the awesome iDEP 0.36, hosted at http://ge-lab.org on Thu Jun 01 10:30:38 2017.
If you really, really need it, here is my embarrassingly messy source code.

Data
Species: Homo sapiens genes (GRCh38.p7)
Number of samples: 6
Number of genes converted and filtered: 13794
2 samples detected.
Input file type: RNA-seq read count file

Pre-processing and exploratory data analysis settings:
Min. counts: minCounts= 10
Counts data transformation method: rlog: regularized log
Method for differential expression: CountsDEGMethod= 3 (DESeq2)
number of genes in heatmap: nGenes= 1000
number of genes in k-means clustering: nGenesKNN= 2000
number of clusters in k-means clustering: nClusters= 2
Promoter analysis for k-means clustering: radioPromoterKmeans= 300 bp

Differential expression settings:
FDR cutoff: limmaPval= 0.1
Fold-change cutoff: limmaFC= 2
Promoter analysis for DEGs: radio.promoter= 300 bp

Pathway analysis settings:
Pathway analysis methods: pathwayMethod= GAGE
FDR cutoff: pathwayPvalCutoff= 0.1
Min size for gene set: minSetSize= 15
Max size for gene set: maxSetSize= 2000

PREDA settings:
FDR cutoff: RegionsPvalCutoff= 0.01
FDR cutoff: StatisticCutoff= 0.5

R session info:

R version 3.4.0 (2017-04-21)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 7 x64 (build 7601) Service Pack 1

… …

Figure 15. User settings and R session information produced by the "R" tab of iDEP.