Wardrobe - an integrated system for analysis of epigenomics and transcriptomics data

Andrey V. Kartashov, M.S.¹ and Artem Barski, Ph.D.¹,²

¹Division of Allergy and Immunology, ²Division of Human Genetics, Cincinnati Children’s Hospital Medical Center and Department of Pediatrics, College of Medicine, University of Cincinnati, Cincinnati, OH

Correspondence: Artem.Barski@cchmc.org
Abstract

High-throughput sequencing has revolutionized biology by enhancing our ability to perform genome-wide studies. However, due to lack of bioinformatics expertise, technologies are still beyond the capabilities of many laboratories. Herein, we present Wardrobe server, which allows users to store, visualize and analyze epigenomics and transcriptomics data using a biologist-friendly GUI without the need for programming expertise. Predefined pipelines allow users to download data from core facilities or public databases, visualize data on a genome browser, calculate RPKMs and identify islands. Advanced capabilities include differential gene expression and binding analysis, and creation of average tag density profiles and heatmaps. Wardrobe code is available at https://code.google.com/p/genome-tools/.

A limited functionality demo-version is available at http://demo.wardrobe.porter.st.

Keywords

Epigenomics, transcriptomics, data analysis, bioinformatics, ChIP-seq, RNA-Seq, DNase-Seq, ATAC-Seq
Introduction

The recent proliferation of next-generation sequencing (NGS)-based methods for analysis of gene expression, chromatin structure and protein-DNA interactions has opened new horizons for molecular biology. These methods include RNA sequencing (RNA-Seq)[1], chromatin immunoprecipitation sequencing (ChIP-Seq)[2], DNase I hypersensitive sites sequencing (DNase-Seq)[3], micrococcal nuclease sequencing (MNase-Seq)[4], global run-on sequencing (GRO-Seq)[5], assay for transposase-accessible chromatin sequencing (ATAC-Seq)[6], and others. On the “wet lab” side, these methods are more or less well established and can be performed by experienced molecular biologists; however, analysis of the resulting data requires computational expertise that many molecular biologists do not possess or have access to. Another difficulty is re-utilization of published datasets: although authors comply with the longstanding requirement to deposit raw data files into databases such as Sequence Read Archive (SRA) or Gene Expression Omnibus (GEO), it is impossible to analyze these datasets without special expertise. Even when processed data files (e.g. gene expression values) are available, direct comparison between datasets is ill advised because different laboratories use different pipelines (or different software versions). This situation means that biologists require the help of bioinformaticians even for simple tasks, such as viewing their own data on a genome browser, and thus places these exciting techniques beyond the reach of the majority of scientists. Even when bioinformaticians are available, differences in priorities can result in delays and misunderstandings that are damaging to the research effort. An optimal way to mitigate these problems is by creating user-friendly tools that can help biologists perform at least some of the tasks without the help of bioinformaticians.

Several stand-alone programs and web services are available for the analysis of NGS data. However, the majority of currently available tools are command line or R based and only a few tools have graphical user interfaces that do not require Linux or programming skills. The programs Genespring [7], Partek [8] and Golden Helix [9] can be run on regular desktop computers and allow analysis of gene expression or
genetic variation. However, users have to load the data manually and store it on their desktop computers; given the sheer volume of NGS datasets, this makes data analysis complicated at best. Furthermore, these tools do not allow for seamless integration of multiple, published or locally produced datasets. Though Illumina Basespace [10] and Galaxy server [11] allow for both storage and analysis of data and have integrated viewing tools. However, they require transfer of data outside the institution (which may be prohibited by HIPAA regulations in some cases) and provide only limited storage space for user data. Although Galaxy provides the opportunity to run analysis without command lines, users still need to select parameters each time, which requires a deep understanding of each tool. Absence of highly formalized pipelines may result in inexperienced users comparing “apples to oranges”. In summary, few of the available tools provide biologist-friendly interfaces, and none integrate data storage, display and analysis with a user-friendly interface.

We therefore developed Wardrobe, a biologist-friendly solution for integrated acquisition, storage, display and analysis of NGS data, primarily for use in epigenomics research. Wardrobe allows for download of raw data from core facilities or databases (e.g. GEO), read mapping and display on an integrated mirror of the University of California, Santa Cruz (UCSC) genome browser [12], quality control and basic and advanced data analysis. The automated basic analysis pipeline incorporates both widely used tools developed by others (e.g. Bowtie[13], STAR [14], FASTX [15] and MACS [16]) and several of our own programs. Customizable advanced analysis includes tools for comparing gene expression DESeq [17] and ChIP-Seq (MAnorm [18]) profiles between samples or groups of samples and creating gene lists, average tag density profiles and heatmaps. Most of the data are precomputed and stored in an SQL database and can be accessed via a convenient web interface by biologists. Bioinformaticians can also access the data via Python or R interfaces on the server or even on the web via third-party tools like RStudio. Wardrobe can be run on Linux or MacOSX systems with attached storage. Source code and

System overview

Wardrobe allows users to upload, store and analyze NGS data. The workflow consists of two parts: basic and advanced analysis (Fig. 1). The basic analysis includes operations that do not require comparison of samples: data download, quality control, calculation of RPKMs (reads per kilobase of transcript per million reads mapped), island identification and upload to a built-in mirror of the UCSC genome browser. Advanced analysis includes comparing gene expression or ChIP-Seq profiles between samples. A flexible data ownership system is implemented: while all users can see all datasets on a local mirror of the UCSC genome browser, only members of the laboratories that own the data can access and analyze datasets within Wardrobe web interface or download it. Laboratory-level administrators can elect to share data with other laboratories. However, trusted bioinformaticians can have access to all datasets outside of web-interface—e.g. via RStudio. We believe that this setup strikes a balance between maintaining data ownership and encouraging collaborations.

Basic analysis

Basic analysis includes operations that are performed on a single library (Fig. 1B). Analysis starts by entering the experiment description into Wardrobe. This information will be used to select the appropriate genome and analysis pipeline. Raw data can be directly downloaded by Wardrobe via hypertext transfer protocol (http) / file transfer protocol (ftp) or from core facilities or internet
databases such as GEO or SRA. Compressed or uncompressed FASTQ (.fastq) and SRA (.sra) files can be used. We elected not to use the prealigned BAM (.bam) files to ensure uniform alignment of samples.

For ChIP-Seq and similar experiments, reads are aligned to the genome with Bowtie [13], quality control analysis is conducted and data are summarized in a table (Fig. 2A). In addition to basic statistics (percentages of mapped/unmapped/non-uniquely mapped reads and average fragment length), Wardrobe displays several other quality control measures. Base frequency plots are used to estimate adapter contamination – a frequent occurrence in low-input ChIP-Seq experiments (Fig. 2B). Average tag density profiles can be used to estimate ChIP enrichment for promoter proximal histone modifications (e.g. histone 3 lysine residue 4 trimethylation [H3K4me3], Fig. 2F). The genome browser can be used to visually compare results to other experiments in the database (Fig. 2C). ChIP-Seq results are displayed on the genome browser as coverage per million reads mapped. For paired-end reads, coverage is calculated as the number of fragments covering each base pair (bp). To obtain coverage for single-read experiments, average fragment length is calculated by MACS [16], and individual reads are extended to this length in the 3’ direction. Islands of enrichment identified by MACS are displayed both on the browser (Fig. 2C) and as a table (Fig. 2D) together with the nearest genes. Use of different parameters or pipelines for different antibodies (e.g. “broad peaks” MACS option for H3K27me3) is possible. The distribution of the islands between genomic areas (promoters, exons, etc.) is displayed as a stacked bar graph (Fig. 2E).

For RNA-Seq analysis, reads are aligned with RNA STAR [14] using the appropriate RefSeq transcriptome. The quality control tab displays the number of reads aligned within and outside the transcriptome. The percentage of the reads mappable to ribosomal (r)DNA is displayed to estimate the quality of rRNA depletion (Fig. 2G). Data are deposited on the browser, and RPKMs are calculated for each transcript (algorithm to be described elsewhere) (Fig. 2H,I). Depending on the application, RPKM values can be
presented for each transcript or summed up for each TSS (for gene expression studies) or for each gene (for functional studies, e.g. Gene Ontology).

**Advanced analysis**

If satisfied with the quality of data obtained from sequencing, a user can proceed to advanced analysis, which involves integration of information from multiple experiments. For gene expression analysis, the typical task is identifying differentially expressed genes. We elected to incorporate the DESeq1/2 algorithm [17, 19] for this purpose because it does not require recreating transcript models and does not make many assumptions. In order to perform gene expression profiling, a user can define replicates and utilize the DESeq algorithm to calculate p-values and fold changes. On the basis of DESeq results, lists of genes whose expression changes can be created within Wardrobe using expression levels, fold change, or p/q-values, as well as other parameters, and downloaded, if needed, in a table form for further analysis (e.g., gene set enrichment analysis) (Fig. 3A).

The gene sets can also be used to create average tag density profiles and heatmaps within Wardrobe (Fig. 3B-F). Average tag density profiles are used to compare the enrichment of histone modifications or other proteins around the TSS or the gene bodies between different gene sets. Often gene bodies used to estimate enrichment, for instance when comparing the levels of positive marks, such as H3K4me3 between expressed and silent genes. Heatmaps provide similar information but allow comparisons of modifications between individual genes. Statistical comparison of tag densities between groups of genes can be performed using Mann-Whitney-Wilcoxon (MWW) tests (Fig. 3C and D). All graphs can be downloaded in publication-quality SVG format.

For ChIP-Seq, the task is usually the identification of areas that have different levels of binding between samples. The difficulty here is that the signal-to-background ratio (enrichment) is usually slightly different between ChIP-Seq experiments; thus, several assumptions have to be made in order to
compare islands of enrichment. Wardrobe uses the MANorm algorithm [18], which assumes that modifications do not change in the majority of areas. This allows MANorm to adjust for differential levels of enrichment between experiments. The lists of islands, fold changes and accompanying p-values are presented in table form, and islands can be viewed in the browser with the push of a button (Fig. 3G).

**Epigenetic changes during Th1 differentiation of human T cells**

To demonstrate the ability of Wardrobe to integrate data obtained from different sources and using different sequencing technologies, we have performed analysis of gene expression and chromatin environment changes during differentiation of human naïve T cells into T helper type 1 (Th1) cells. For this purpose, we utilized a gene expression and chromatin dataset published by Hawkins et al. (SRA082670) [20]. The study included RNA-Seq data for resting naïve T cells (“T helper precursor cells [Thp]” in the paper) and T cells activated in Th1 conditions for 72 hours in triplicate and ChIP-Seq data for T cells activated in Th1 conditions. We wanted to identify chromatin changes that occur during Th1 differentiation of T cells, but the Hawkins et al. study did not include chromatin profiles for resting naïve cells. For this reason, we used H3K4me3 ChIP-Seq data for resting naïve T cells obtained for a separate project in our laboratory. RNA-Seq data was obtained using a Helicos platform, whereas ChIP-seq libraries were sequenced on Illumina.

After entering the sample descriptions and links to .sra files from the SRA database into the system, Wardrobe downloaded the dataset and performed basic analysis (Fig. 2A). ChIP-Seq data demonstrated the expected percentage of reads mapped (Fig. 2B), average tag density profiles showed high enrichment at promoters (Fig. 2F), MACS identified ~48 thousand and 79 thousand islands (naïve T cells and Th1 cells, respectively, Fig. 2D), the majority of which (68-77%) were located at promoters (Fig. 2E). However, RNA-Seq results demonstrated poor mapping to the human transcriptome, especially for naïve T cell samples (2-6%) (Fig. 2G), whereas 35% of reads were mapped outside the annotated
transcriptome. Altogether, these quality control measurements led us to conclude that the RNA-Seq libraries were severely contaminated by genomic DNA and that the RPKM values are likely to be inflated by several RPKMs, in particular in the naïve samples. Nonetheless, we proceeded with the advanced analysis.

We next performed comparison of gene expression using DESeq2. Replicates were defined, genes were grouped by common TSS and differentially expressed genes were identified. DESeq results (Fig. 3A) were used to define lists of genes that were expressed or silent in both Thp and Th1 cells or that were induced during differentiation. These lists were used to create average tag density profiles in both naïve and Th1 cells (Fig. 3B,D,E). As demonstrated in the graphs and MWW statistical analysis (Fig. 3D), expressed genes have higher levels of H3K4me3 at their promoters than the silent genes. Inducible genes had intermediate levels of this modification in naïve cells, in which these genes were silent, suggesting that H3K4me3 poises inducible genes for expression. The same conclusion can be made upon examining tag density heatmaps (Fig. 3B). Upon induction, the H3K4me3 levels of the genes in the induced list increased to the level of the genes in the expressed list.

The H3K4me3 ChIP-Seq experiments worked well in both naïve and Th1 cells; however, if we compare the height of the TSS peaks of H3K4me3 between naïve and Th1 cells (Fig. 2D and data not shown), we can conclude that the enrichment in the naïve experiment is slightly higher. This difference in enrichment would affect the peak calling, meaning that the peaks cannot be compared directly between the two cell types. To overcome this obstacle, Wardrobe uses the MAnorm algorithm, which compares the read numbers between islands while adjusting for enrichment [18]. MAnorm produces fold changes and p-values for the peaks, and the areas of interest can be viewed on the browser (Fig. 3F).
Implementation

Wardrobe is accessible via Google Chrome, Safari and Firefox browsers. The user interface is web based and utilizes HTML5 and JavaScript technologies. To speed up the development process, EXTJS and D3 frameworks were used. On the server, Apache with PHP is used to process user’s requests. Linux or MacOSX native job schedulers are used to run Python pipelines. For stability, all pipelines have separate queues and process statuses. Pipeline output is stored in the SQL database with the exception of BAM files. These precomputed data are accessible by third-party software like RStudio that allows analysis that is not included in Wardrobe. There are no specific hardware limitations for Wardrobe. We have installed it on both a Linux server and Mac Pro desktop and laptop computers. An average Intel Core i7 computer with 32 gigabytes of RAM and a SATA HDD; more than 100 M read/write speed) will analyze one ChIP-Seq or RNA-Seq experiment within 2 hours.


In summary, we have developed an automated system for storage and analysis of NGS data. The Wardrobe system can be easily installed on Mac or Linux computers and provides data analysis solutions for the whole laboratory or institution.
List of abbreviations used

ATAC-Seq  assay for transposase-accessible chromatin sequencing

ChIP  chromatin immunoprecipitation

ChIP-Seq  chromatin immunoprecipitation sequencing

D3  data-driven documents

DNase-Seq  DNase I hypersensitive sites sequencing

Ext JS  extension JavaScript

ftp  file transfer protocol

GEO  Gene Expression Omnibus

GRO-Seq  global run-on sequencing

HIPAA  Health Insurance Portability and Accountability Act of 1996

HTML  hypertext markup language

MACS  model-based analysis of ChIP-Seq

MAnorm  a model for quantitative comparison of ChIP-Seq datasets

Me3  trimethylation

MNase-Seq  micrococcal nuclease sequencing

MWW  Mann-Whitney-Wilcoxon

NGS  next-generation sequencing
PHP | PHP: hypertext preprocessor
---|---
RAM | random-access memory
RefSeq | National Center for Biotechnology Information Reference Sequence
RPKMs | reads per kilobase of transcript per million reads mapped
RNA-Seq | ribonucleic acid sequencing
SATA HDD | serial ATA hard disk drive
SRA | Sequence Read Archive (formerly known as Short Read Archive)
STAR | spliced transcripts alignment to a reference
SVG | scalable vector graphics
TSS | transcription start site(s)
Th1 | T helper type 1
Thp | T helper precursor cells
UCSC | University of California, Santa Cruz

**Competing interests**

The authors declare no competing financial interests.

**Authors’ contributions**

AK and AB designed the system, AK wrote the program, AK and AB wrote the paper.

**Authors’ information**

Andrey V. Kartashov, M.S.¹ and Artem Barski, Ph.D.¹,²
Acknowledgements

Authors would like to thank Shawna Hottinger (CCHMC) for editorial assistance.

The work was supported in part by the Children’s Hospital Research Foundation and NHLBI NIH career transition award to A.B. (K22 HL098691).
References


8. Partek [www.partek.com]


10. Illumina basespace [https://basespace.illumina.com/home/index]


15. FASTX [http://hannonlab.cshl.edu/fastx_toolkit/index.html]


Figure legends

**Figure 1. Wardrobe overview and pipelines.** A. The Wardrobe server can be set up on a Linux or Mac computer attached to a consumer-class storage array, typically within a local institutional network. Biologists can use Wardrobe to upload data from a sequencing core or a public database and promptly receive quality control data, view the results in the browser and perform some of the analysis without the assistance of bioinformaticians. Bioinformaticians can access the precomputed data in Wardrobe’s SQL database to perform further analysis. B. Basic analysis pipelines. The flow diagram shows the tools used in the basic analysis pipelines for RNA-Seq and ChIP-Seq data. C. Advanced analysis allows the user to identify differentially expressed genes using DESeq, create gene lists and use these list to generate average tag density profiles and heatmaps. Differentially bound areas can be identified with MAnorm.

**Figure 2. Basic analysis.** A. The laboratory data table shows the list of experiments available to the user. B. Shown is the quality control tab for a ChIP-Seq experiment; the spiky base frequency plot indicates minor adapter contamination. C. Shown is the ChIP-Seq browser shot for the CD4 gene, visible on the genome browser tab. D. The island list tab shows locations of islands and the nearest genes. E. The island distribution bar graph, visible from the islands distribution tab, shows that the majority of H3K4me3 islands are located at promoters. F. The average tag density profile, visible from the average tag density tab, shows enrichment around the TSS. G. Shown is the quality control tab for an RNA-Seq experiment. The base frequency plot shows an AT bias, suggesting DNA contamination and read-length variation characteristic of Helicos sequencing. H. Shown is the RPKM table available from the RPKM list tab for an RNA-Seq experiment. I. Shown is the RNA-Seq browser for the CD4 gene, visible from the genome browser tab.

**Figure 3. Advanced analysis.** A. The table displays the results of DESeq2 differential gene expression analysis, showing induction of TNF upon Th1 differentiation. B. The heatmaps show expression and
H3K4me3 tag density in the three examined sets of genes (Silent, Expressed, Induced) in naïve T cells and Th1-differentiated T cells.  

C, E, F Shown are the H3K4me3 average tag density profiles for the gene body (C) or around the TSS (E, F) in naïve T cells and Th1-differentiated T cells.  

D. The box plot shows the distribution of H3K4me3 tag densities for the three gene sets within the area shaded in panel C (Silent, Expressed, Induced). MWW p-values are shown below the box plot.  

G. Shown is identification of differentially modified areas using MAnorm. Notice the appearance of H3K4me3 islands at Th1 characteristic cytokine IFNG.
**Kartashov Figure 3**

### A DESeq

<table>
<thead>
<tr>
<th>refseq_id</th>
<th>gene_id</th>
<th>RPKM_Naive</th>
<th>RPKM_Th1</th>
<th>LOGR</th>
<th>pvalue</th>
<th>padj</th>
</tr>
</thead>
<tbody>
<tr>
<td>NM_002160</td>
<td>TNC</td>
<td>0.299162</td>
<td>0.200689</td>
<td>-2.259</td>
<td>1.06E-05</td>
<td>6.81E-05</td>
</tr>
<tr>
<td>NM_000594</td>
<td>TNF</td>
<td>0.614744663</td>
<td>36.1460573</td>
<td>3.84</td>
<td>9.74E-35</td>
<td>9.14E-33</td>
</tr>
<tr>
<td>NM_021137</td>
<td>TNFA/F1</td>
<td>0.437074671</td>
<td>6.5483</td>
<td>0.35375</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### B Heatmaps

- **Expression**
  - Naive
  - Th1

- **Modifications**
  - Naive
  - Th1

### C Tag Density Profiles

- **MWW p-values**
  - Silent
  - Expressed
  - Induced

### D Average Tag Density Plots

**H3K4me3 in Naive**

- **H3K4me3 in Th1**

### E Gene Body Average Tag Density

**Tag Density Profiles**

- **Induced**
- **Expressed**
- **Silent**

### F DESeq

<table>
<thead>
<tr>
<th>refseq_id</th>
<th>gene_id</th>
<th>strand</th>
<th>chrom</th>
<th>start</th>
<th>end</th>
<th>description</th>
<th>raw_read1</th>
<th>raw_read2</th>
<th>H_value_scaled</th>
<th>A_value_scaled</th>
<th>log10_p_value</th>
<th>region</th>
</tr>
</thead>
<tbody>
<tr>
<td>NM_002160</td>
<td>TNC</td>
<td>-</td>
<td>chr12</td>
<td>60954925</td>
<td>6095721</td>
<td>unique_peak1</td>
<td>11</td>
<td>0</td>
<td>2.591239412469</td>
<td>1.289405767194</td>
<td>2.072092328326</td>
<td>intergenic</td>
</tr>
<tr>
<td>NM_000594</td>
<td>TNF</td>
<td>-</td>
<td>chr12</td>
<td>609532680</td>
<td>609532994</td>
<td>unique_peak1</td>
<td>11</td>
<td>0</td>
<td>2.591239412469</td>
<td>1.289405767194</td>
<td>2.072092328326</td>
<td>intergenic</td>
</tr>
<tr>
<td>NM_006619</td>
<td>IFNG</td>
<td>-</td>
<td>chr12</td>
<td>60546325</td>
<td>60546373</td>
<td>unique_peak1</td>
<td>11</td>
<td>0</td>
<td>2.591239412469</td>
<td>1.289405767194</td>
<td>2.072092328326</td>
<td>intergenic</td>
</tr>
<tr>
<td>NM_015334</td>
<td>IFNGR2</td>
<td>+</td>
<td>chr21</td>
<td>54773752</td>
<td>54746366</td>
<td>merged_peak</td>
<td>73</td>
<td>0</td>
<td>2.591239412469</td>
<td>1.289405767194</td>
<td>2.072092328326</td>
<td>intergenic</td>
</tr>
</tbody>
</table>