Massive growth in the amount of research data in biology has led to increasing need for user-friendly analysis tools. We developed BioWardrobe platform to simplify routine analysis of epigenomics data. It had scheduled python scripts as a backend and a web interface to display the results. However, growing utilization of BioWardrobe revealed that the python back-end lacked functionality: it was difficult to add new or update existing pipelines and it could not optimally utilize resources. The recent development of Common Workflow Language (CWL)\(^1\) allowed us to solve these challenges. Thus, we combined CWL with an open-source pipeline manager, Airflow, to create a new back-end. BioWardrobe’s pipelines were rewritten in CWL, component tools were containerized with Docker, and Airflow was extended with CWL support. This enabled BioWardrobe to have more control over pipeline execution, which in turn helped to save time and reduce computational costs. CWL-based workflows are compact, well formalized, and can be easily versioned and visually represented making them more scientist friendly. Utilizing CWL-Airflow allows to set up analysis in any hardware environment from stand-alone server to cluster or cloud. The simplified installation of CWL-Airflow and pipelines enables scientists to concentrate on analyzing the data. BioWardrobe can be found at [http://biowardrobe.com](http://biowardrobe.com).

Figure 1. BioWardrobe 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 list 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 2. BioWardrobe pipelines in CWL Airflow. Old pipelines for processing RNA-Seq and ChIP-Seq data were converted into CWL. Furthermore, BioWardrobe functionality was extended with several new pipelines. The same setup on Mac and Linux servers is used, with the same tools and versions, thanks to Docker. That leads to homogeneous workflow environment in development and production.

Figure 3. A workflow in CWL Airflow. Airflow’s DAG class was extended with ability to parse CWL workflows, CWL steps were converted into Airflow’s operators and organized into Airflow’s DAG.

Figure 4. CWL Airflow. CWL Airflow supports all the features that are needed: status for each step can be checked, workflow can be restarted from a specific step and the log file can be checked. The stalled workflows can be easily found and the failed steps can be resolved.

Figure 4. Rabix Composer. Graphical workflow assembly is possible with tools like Rabix Composer.

References