Novel in silico end-to-end platform tested on mesenchymal stromal cells (hMSC) revealed dysregulated MAB21L2 and CXCL6 in primary osteoporosis

Catherine Jessica Lai

Brearley School, New York, NY 10021, USA

ABSTRACT

The dramatic drop in genome sequencing costs has led to vast stores of data, but the fragmentation of bioinformatic tools to analyze the data continues to elude many biologists. Mendel is a new R-based platform that allows biologists to focus on biology without the burden of learning software engineering. It is the first system to provide users the full range of an end-to-end experience, starting from uploading sequencing data through full-fledged differential analyses and functional annotation. Mendel has been tested on the transcriptome of human mesenchymal stromal cells (hMSC), using Affymetrix U133 Plus 2.0 RNA probe data from 5 primary osteoporosis and 5 control subjects. Importantly, the test also enhanced our current understanding of the disease, revealing additional and novel over-expressed genes MAB21L2 and XIST, and discovering a new series of downregulated genes CXCL6 and CMPK2. Mendel also reveals pathways, such as that for COL1A1/COL10A1, whose upregulation delays post-translation heterotrimer folding, leading to abnormal α-chain collagens and helices, the latter of which are either degraded by the ERAD pathway or secreted into the ECM, leading to matrix mineralization, osteoblast development, and cell-cell crosstalk.

Keywords: Bioinformatics, differential gene analysis, expression profiling, gene ontology, osteoporosis

INTRODUCTION

The cost for sequencing the human genome has plunged from US$10 million to a few thousand dollars in the last six years [5]. This drop has led to vast stores of data, but the fragmentation of bioinformatic tools to analyze the data continues to elude many biologists. The number of packages written in the popular R language has grown exponentially to over 3,500 by 2012 [6]. Worse, these packages work on different platforms, use a bewildering variety of data formats, employ incompatible user interfaces, undertake narrowly-defined analyses with tool-specific assumptions, and yet do not have application programming interfaces (APIs) to connect themselves into a coherent platform. Therefore, contemporary biologists often spend months to undertake the key tasks of pre-processing, differential gene analyses, and functional annotation.

Recently, a consensus is emerging on the tasks required to analyze genomic data. Mendel is a new R-based platform that exploits this emerging template for data analyses, so that biologists can focus on biology without the burden of learning software engineering. It is the first system that allows users the full range of an end-to-end experience, starting from uploading sequencing data through full-fledged differential analyses. Mendel has been used tested on the transcriptome of human mesenchymal stromal cells (hMSC), using Affymetrix U133 Plus 2.0 RNA probe data from 5 primary osteoporosis and 5 control subjects. Importantly, the test enhanced our current understanding of the disease, revealing additional and novel over-expressed genes MAB21L2 and XIST, and discovering a new series of downregulated genes CXCL6 and CMPK2.
In the “Materials and Methods” section, we describe Mendel’s architecture and functionality. We also describe how we test Mendel using the hMSC cell lines. Then we report the results of the test, and conclude with thoughts on some remaining limitations and further research directions.

MATERIALS AND METHODS

Mendel is developed using the R language. There are several advantages to using R. First, it has powerful statistical functions that facilitate the computations involved in bioinformatics analyses, such as single-channel regression design models. Second, and as a consequence of the first, R is popular among biologists who do programming. This popularity feeds into a third factor, which is that many coders created function-specific packages in R. This is both a boon—many sophisticated functions are now available—and a bane—each package is often developed without consideration given to other packages. This fragmentation motivates our development of R.

R is supported on most operating systems. Mendel is developed on Windows 8 using the RStudio development environment. Figure 1 shows its architecture. Mendel uses Shiny interface components, and a curated set of R packages show in the figure. The entire system is open-sourced, and we have made Mendel freely available at http://mendel.crabdance.com.

We selected packages based on the functionality now common among recent research papers on differential gene analysis and expression profiling [1, 2, 7, 9, 13, 14]. In particular, the following functions are incorporated:

- Pre-processing with Robust Multichip Average (RMA) analysis. This achieves three goals: (1) background correction to enhance the array’s sensitivity by adjusting the intensity readings for non-specific signals, (2) between-array normalization to account for differences in handling, labeling, hybridization, and (3) scanning, and third, reporter summarization to compute expression values for each gene using the array probes that target the gene’s transcripts.
- Differentially expressed gene (DEG) analysis. Mendel uses a single-channel design that provides for standard treatment-contrast parameterization, which has a coefficient for the mutant versus wild-type difference.
• Gene ontology (GO) annotation. Mendel provides all three standard levels of analyses: (1) biological process (BP), (2) molecular function (MF), and (3) cellular components (CC). Standard GO uses the popular hypergeometric model, which finds genes with large differences, but it can miss genes that work in a coordinated way in a set of related genes. So Mendel also provides for gene set enrichment analysis (GSEA). This determines whether an a priori set of genes has significant concordant differences between the phenotypes of diseased and control.

• Disease ontology (DO) annotation. This complements GO, and displays disease associations of our discriminating genes. Mendel also provides GSEA variants of DO.

• Pathway analyses. Mendel displays pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG).

We conducted many tests using Mendel, and report one such test using hMSC cell lines from the GSE35936 from NCBI [1]. The dataset consisting of 1,354,896 million probes for 54,675 genes in each of 10 cell lines, of which 5 are age-matched controls. The data was obtained using the Affymetrix U133 Plus 2.0 Array platform.

RESULTS AND DISCUSSION

The first key finding is that Mendel has been successfully implemented in a way that biologists can undertake without being burdened by software complications. The second finding is that Mendel is validated using a replication study. It is able to report the same findings as in Benisch, et al, [1]. Third, Mendel could even enhance the findings to report novel results beyond Benisch, et al.

Figure 2 reports the distribution of intensities for one cell line in our test dataset. It shows that leftmost five control chips look more similar to each other in terms of means and ranges, than the rightmost five diseased chips. This is the first indication that a DEG might be fruitful.
Figure 3 shows a heatmap, which gives a gene-level confirmation of the observation that the control and diseased cell lines likely have different intensities.

The differentially expressed genes are shown in Figure 4, for upregulated genes. Benisch, et al. [1] reported that COL1A1/COL1OA1 and MAB21L2 are the most upregulated. Mendel’s analysis also reports the same, validating the analysis. Figure 5 shows that the upregulated genes are strongly correlated. A regression of the Mendel log fold-changes on the Benisch log fold-changes has a p-value of 0.000025.

In addition, Mendel discovers some significant downregulated genes, such as CXCL6 and CMPK2 (Figure 6). The log fold-changes for these are even larger than those for COL1A1/COL1OA1 and MAB21L2, suggesting that these downregulated genes may be good candidates for therapeutic targeting.
Figure 4 - Differentially expressed genes (upregulated).
Figure 5 - Correlation between Mendel and Benisch, et al. log FC for upregulated genes
To get a better grip on the biology behind these differentially expressed genes, Mendel provides a gene ontology annotation of the dataset. Figure 7(a) reports the result for GO from a BP perspective, and in particular, Figure 7(b) shows the “regulation of transcription from RNA polymerase II promoter,” with a GSEA enrichment score of 0.205. Mendel can also produce visualizations of all the ontologies in a gene map or histogram (see menu in Figure 7(a)).

Figure 8 shows the result of the disease ontology annotation, which identifies associations with cell adhesion molecules (CAMs), ubiquitin mediated proteolysis, among others. As with GO annotation, Mendel provides gene-disease maps of these DOs (see menu in Figure 8).
(a) Gene ontology with BP, using GSEA.

(b) “Regulation of transcription from RNA polymerase II promoter.”

Figure 7 - Gene ontology annotation
Figure 9(a) reports the results of pathway analyses, and Figure 9(b) shows what a biologist can see by selecting the top-ranked pathway: ECM-receptor interactions. Mendel also allows the user to visually see the pathways individually and as an interconnected set.

Through annotations such as GO, DO, and pathway identification, a biologist can easily and quickly size up the mechanism behind the differentially expressed genes. For example, one of the interesting results of our osteoporosis test is that it reveals significant upregulated COL1A1/COL10A1. The annotations allow us to map out the biology of how COL1A1/COL10A1 could have led to primary osteoporosis (Figure 10).
(a) Pathway identification in Mendel
(b) ECM-receptor interaction pathway

Figure 9 - Pathway identification
CONCLUSION

The analysis of enormous amounts of genomic data is hampered by the current fragmentation of tools. Biologists often take months to learn how to use these tools, determine which ones work best, and which ones work together as a whole. Mendel provides a platform that integrates the key tools that can provide the analytics used in the recent literature. It is validated with a test using live data on primary osteoporosis. Not only does Mendel provide findings similar to published results, it even provides additional novel and significant findings, such as downregulated differentially expressed genes.

Going forward, Mendel can be improved by with additional functionality, but we urge caution in doing so, since greater functionality may confuse the biology user more than it benefits her. Another path for future enhancements is to open Mendel to power users who might be able to enhance Mendel’s functionality on their own. This may ultimately be the wiser move, since it allows only those who want or can extend Mendel to do so.

Acknowledgements
I thank Prof. P. Schlar of Mount Sinai Hospital for guiding me towards R and to Peggy Benisch at the University of Wuerzburg and her colleagues for usage of the primary osteoporosis dataset. I am also grateful to the developers [3, 4, 8, 10, 11, 12, 15, 16, 17, 18] of the R packages used in Mendel.
REFERENCES