Konrad-Zuse-Zentrum
für Informationstechnik Berlin

Takustraße 7
D-14195 Berlin-Dahlem
Germany

XINTIAN ARTHUR YOU[1]

Hastings: An R pipeline for large-scale RNA-Seq data analysis

[1] *Zuse Institute Berlin, Germany*

ZIB Report 15–40 (September 2015)

Herausgegeben vom
Konrad-Zuse-Zentrum für Informationstechnik Berlin
Takustraße 7
D-14195 Berlin-Dahlem

Telefon: 030-84185-0
Telefax: 030-84185-125

e-mail: bibliothek@zib.de
URL: http://www.zib.de

ZIB-Report (Print) ISSN 0000-0000
ZIB-Report (Internet) ISSN 0000-0000

Abstract

Motivation: With the advance of high-throughput sequencing technologies, large-scale datasets becomes increasingly common, from thousands of patients involved in a cohort study to thousands of single cells from the same tissue. Despite major advances in understanding the molecular mechanisms governing biological processes from development to diseases, heterogeneity from between two individual persons to two cells from the same tissue is yet to be scrutinized. Several studies have revealed new insights from such large-scale datasets [1, 2], yet their analysis protocols are not always straightforward to be adopted or extend. In particular, it is often more of an art to select various parameters along the whole analysis procedures, and visualization of high dimensional data could be very challenging.
Results: Here, we present Hastings to face the demand of large-scale data analysis and visualization for RNA-Seq gene expression data. As demonstrated in the three examples, Hastings can efficiently identify sub-groups in an unsupervised manner, identify potential marker genes and generate clear 2D visualizations. Hastings could be widely applied from bench to clinics.
Availability: plan to go on Github

1 Introduction

Each Leonardo Da Vinci's egg is different from the others, and likewise for human beings, tissues and individual cells in terms of genotype and phenotype. To date, most biological studies compare two or more groups/conditions in order to learn regulatory mechanisms from the inter-sample heterogeneity. Only very recently, the importance of intra-sample heterogeneity is appreciated thanks to the advance in the high-throughput sequencing (HTS) technology and the single-cell capture techniques. With the availability of large-scale data (gene expression from thousands of people (TCGA) [1] or thousands of cells from one tissue [3]), new biological insights can be revealed from the aspect of intra-sample heterogeneity. However, most state-of-the-art analysis tools for HTS take in thousands of measurements (such as abundance of genes) but only few samples (such as two conditions each with triplicates), and they normally require prior knowledge of the conditions and might be difficult to extend to handle large-scale datasets. Since there is a shift in the analysis paradigm, namely from inter-sample difference to intra-sample difference together with a demand of unsupervised sub-group classification, new analysis tools are in need. Here, we present Hastings as a user-friendly R pipeline that can process large-scale RNA-Seq data, perform unsupervised classification, build phylogenetic tree, identify markers, and visualize heterogeneities between and within samples.

2 Methods

The pipeline of Hastings is depicted in Figure 1 :

First, low quality data points are removed and a set of highly variable genes (HVG) are selected for further analysis. Second, Singular Value Decomposition (SVD) is performed on HVG data, in conjunction with permutation tests to select only a few significant principal components (sigPC) that can explain true variance among samples other than noise. In this step, the fine structure of the dataset is identified and the noise most likely introduced in the experimental procedures is largely removed. Third, the dimension is further reduced to 2 using t-distributed stochastic neighbour embedding (t-SNE) [4] and a phylogenetic tree is

drawn using Mahalanobis distance. The advantage of this approach is to avoid the rather arbitrary selection of the cluster numbers as in k-means clustering or the the distance threshold in the hierarchical clustering. Also, the dataset is structured in a way that the relationship among each data points is maintained and visualized in 2D scatter plot, allowing for further manual inspections and re-analysis. Finally, marker genes of different clusters are identified and visualized.

HVG can be optionally adjusted after step two, by choosing the genes with top loading values after projecting all genes onto the sigPCs. Then a second round of SVD is performed using this new gene set and the new set of principal components are determined accordingly. The markers identified can be used for further investigation and even clinical diagnosis. Alternatively, Hastings can be used in a "review" mode, in which a pre-defined marker set in used as HVGs.

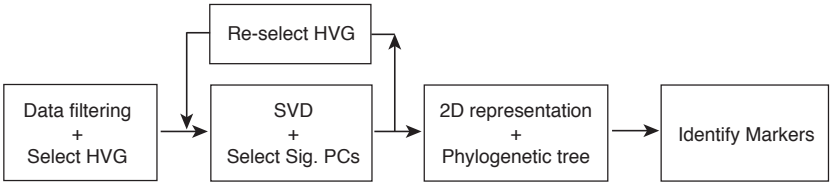

Figure 1: Workflow of Hastings.

2.1 Data filtering and HVG selection

To remove noise introduced in the experimental procedures, customised filtering criteria should be applied prior to further analysis. Typical filtering criteria includes: 1) remove measurement (such as gene or transcript isoform) entries when the value is below certain threshold (e.g. $TPM < 1$); 2) remove sample entries when the confident measurements is below certain threshold (e.g. less than 1000 genes expressed). Appropriate normalization procedures such as quantile normalization or L^1-regularization could be applied.

To focus on the informative measurements as well as to speed up the analysis, a subset of genes that most likely to explain the majority of the heterogeneity is selected. Intuitively, these genes should have high variance among all the data points. Therefore, HVGs are selected by choosing the top 10% of the expressed genes in terms of either Fano factor or CV. Alternatively, when Hastings is used in a "review" mode, a pre-defined marker set can by used instead.

2.2 Singular Value Decomposition

Let X be an $m \times n$ matrix, rows for genes and columns for samples, and the SVD is:

$$X = U\Sigma V^T$$

where Σ is a $k \times k, k \leqslant min(m, n)$ diagonal matrix with non-negative real numbers on the diagonal (the i^{th} values is proportional to the square root of the variance explained by the i^{th} PC). U is an $m \times k$ unitary matrix that projects the original X in k dimensional space focusing on the gene, and V^T is a $k \times n$ unitary matrix that project X in another k dimensions focusing on the samples. Finding a smallest k that keeps as much information as possible is the key in large-scale data analysis and it can be done using permutation test as described in the next section.

The Singular Value Decomposition can be done using an R package "svd". When dealing with large-scale datasets, the dimension could be so high that it would take svd hours or even days to finish. Therefore, randomized SVD (RedSVD [5] or flashpca [6]) can be used instead with marginal loss of accuracy. Hastings makes a test run using all three method and selects the best one for permutation tests.

2.3 Selecting Informative PCs

Permutation tests are used to select principal components (PC) that explain true variance other than noise. The rationale is that a cluster-discriminative gene should have a large weight on a projected PC, whilst shuffling its expression among samples should decrease the capability of cluster separation and thereby decreases the projection weight of this gene [3]. For each permutation, the expression of a small proportion (e.g. 1%, so that the PCs remains largely the same) of the genes are shuffled among samples prior to SVD, and the projection weight of these genes are stored. After permutations, the weights of new projections are compared with the originals for each PC, and a empirical p-value is derived. PCs with p-values below a certain threshold are selected.

2.4 Two-dimensional representation

When there are more than 2 dimensions after permutation test, t-SNE is applied to generate 2D representation. In this representation, the euclidean distance between samples is similar to that in the original high-dimensional space. And a subsequent classification can easily be illustrated using Density-based spatial clustering of applications with noise (DBSCAN) [7]. Hastings optimizes several important parameters (such as perplexity and epsilon-distance) by minimizing the sum of squared error (SSE) of each trail and its perturbations.

2.5 Phylogenetic analysis

A phylogenetic tree could be constructed based on dissimilarity metric using an R package "hclust". However, due to the high dimensionality of the large-scale or single-cell experiments together with the heterogeneity among samples and noise in measurement, it is advisable to use the Mahalanobis or Jensen-Shannon distance on the data representation on the reduced dimensions instead of the commonly used euclidean distance on the raw expression matrix. Marker genes for one cluster or a set of closely related clusters can be subsequently identified using statistical tests such as t-test.

3 Results

To demonstrate the versatile application potential of Hastings, we tested it on three large-scale datasets each with different biological interest. On TCGA dataset, Hastings separated with high accuracy different cancer types using unsupervised classification method. On DRG dataset, Hastings identified many sub-cell types with distinct marker genes from hundreds of DRG neurons that was though to consist of three cell types. On CTC dataset, Hastings separated CTCs from cell-lines and primary tumor samples and revealed substantial heterogeneity for cells derived from the same person, which emphasized the importance of single-cell analysis in, both early-stage and late-stage, cancer diagnosis.

3.1 TCGA dataset

Molecular markers are important for both clinical diagnosis and a better understanding of the underlying mechanism of various diseases, including cancer. Application of Hastings to a dateset from an RNA-Seq experiment on 3602 patients of 12 cancer types [1] demonstrated that cancer types can be clearly separated. Patients were classified into 10 groups (Figure 2A), with group-8 corresponding to colon (COAD) and rectal (READ) adenocarcinomas, and group-4 corresponding to head and neck squamous cell carcinoma (HNSC) and lung squamous cell carcinoma (LUSC). This finding is consistent with the original study that COAD and READ are grouped together presumably due the similarity of tissue of origin, and that LUSC and HNSC are grouped together presumably due to the similarity in gene expression pattern that underlies the common squamous morphology. The unsupervised classification achieves 97.08% accuracy comparing to the histological examination, higher than the result in the original study (84.38% using only mRNA data, and 89.56% from cluster-of-cluster assignments using six data sources). A phylogenetic tree of the classified groups is shown in Figure 2B and hundreds of marker genes are identified and visualized in Figure 2C. Interestingly, although LUSC and HNSC samples are grouped together as previously suggested, they could still be separeted as LUSC localized to the left half of the group whereas NHSC localized to the right (Figure 2A). This separation is more evident by different expression pattern of marker genes in Figure 2C. Marker genes identified here can serve as clinical markers for cancer type diagnosis.

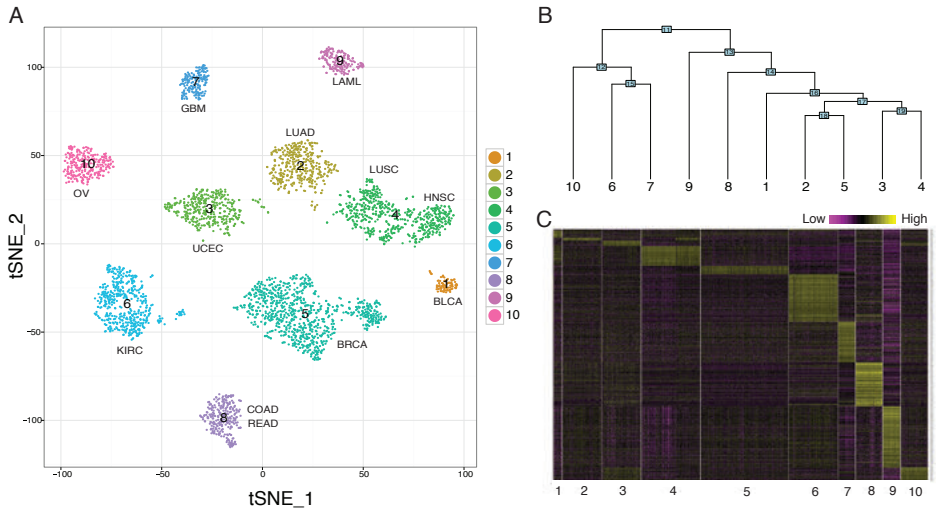

Figure 2: TCGA dataset. A. Unsupervised classification of 3602 patients. Overall, each group represent one distinct cancer type, excepted that group-8 correspondes to colon and rectal cancer, and group-4 correspondes to two squamous cancers LUSC and HNSC. B. A phylogenetic tree among groups. C. Expression pattern of marker genes for all groups.

3.2 DRG dataset

After demonstrating that gene expression can be used to separate different cancer types, or more broadly speaking, different tissues, we went on the identify sub-types of cells from the same tissue. Here, we applied Hastings to an exploratory dataset from a Single-Cell

RNA-Seq experiment on 622 mouse dorsal root ganglion (DRG) neurons [8]. Neurons are classified into 13 groups (Figure 3A). And when examining expression pattern of the known markers for a few well studied cell types (NF, NP, PEP and TH), each of the known cell types contains a few clearly separable subgroups (Figure 3A). A phylogenetic tree is shown in Figure 3B and marker genes are identified and visualized in Figure 3C. Owing to the nature of Single-Cell RNA sequencing, the expression noise between cell from the same group could not only originate from the experimental protocol but also reflect the important information of cell-to-cell heterogeneity.

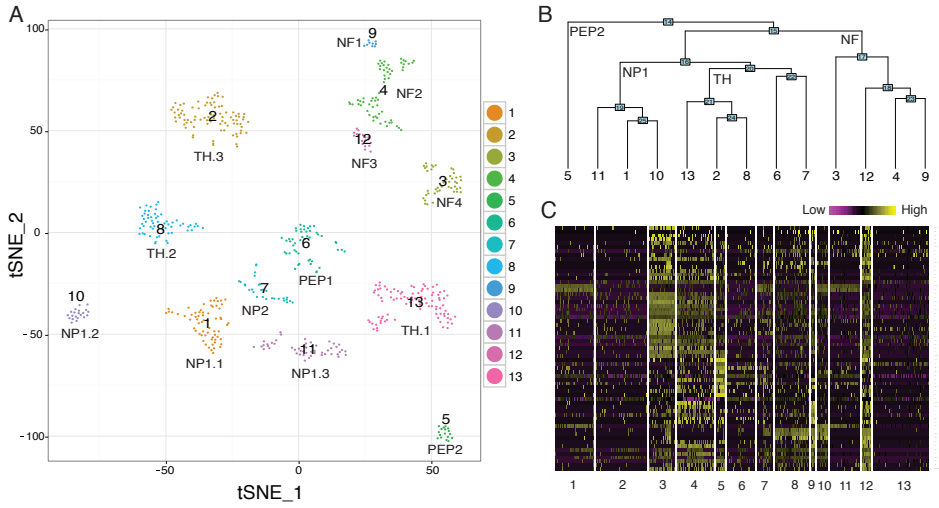

Figure 3: DRG dataset. A. Unsupervised classification of 622 mouse DRG neurons. Overall, four known major types of DRG neurons (NF, NP, PEP and TH) are well separeted and each major type can be further divided into distinct sub-groups. B. A phylogenetic tree among groups. C. Expression pattern of marker genes for all groups.

3.3 CTC dataset

The knowledge of heterogeneity is particularly important for personalized treatment of cancer. For instance, prostate cancer in initially responsive to androgen deprivation, but the efficacy of the androgen receptor (AR) inhibitors varies in later stages. Whilst traditional biopsy is challenging, liquid biopsy by sampling circulating tumor cells (CTC) may reveal drug-resistance mechanism [9]. To demonstrate the potential of Hastings to be applied in medical diagnosis, we applied it to a dataset from this Single-Cell RNA-Seq experiment consisting of 124 CTCs from 22 patients,30 cells from 4 cell lines (VCaP, LNCaP, PC3, and DU145), three patient-derived leukocyte controls and bulk RNA-Seq experiment on primary prostate cancers from 12 patients. To demonstrate the strength of noise reduction capacity of Hastings, we used all the sequencing samples instead of selecting a subset of CTCs using an arbitrary threshold as in the original study. Samples are classified into nine groups, and primary tumors (group-9), cell lines (group-8) and CTCs (majorly group-1,2,7) are clearly separated (Figure 4A). It is now more obvious that single CTCs from each individual demonstrated considerable heterogeneity (Figure 4B) than the original heatmap [9, Fig.1A]. There is clearly a transition from early-stage/enzalutamide-naive CTC (group-1) to later-stage/enzalutamide-treated CTCs (group-2). Group-7 largely constitutes of circulating cells whose lineage cannot be confirmed by canonical marker gene analysis, and their gene expression pattern is indeed

different from CTCs, healty white blood cells or prostate cell lines.. Group-8 consists of all four cell lines, with two cell lines (VCap and LNCap, AR sensitive and lowly tumorigenic) positioned closer to all other single cells comparing to the other two cell lines (DU145 and PC3, AR insensitive and moderately-to-highly tumorigenic). CTCs display moderate-to-high expression of epithelial marker (Figure 4C), whereas the mesenchymal marker is not expressed compared to primary tumors (Figure 4D). A subset of CTCs express MYC and WNT7B (a non-canonical Wnt ligand), which might provide survival signals facing AR inhibition (Figure 4E,4F). Taken together, these findings suggest the importance of single-cell study for the mechanisms underlying cancer development and even might help to design suitable treatment for each individual patients.

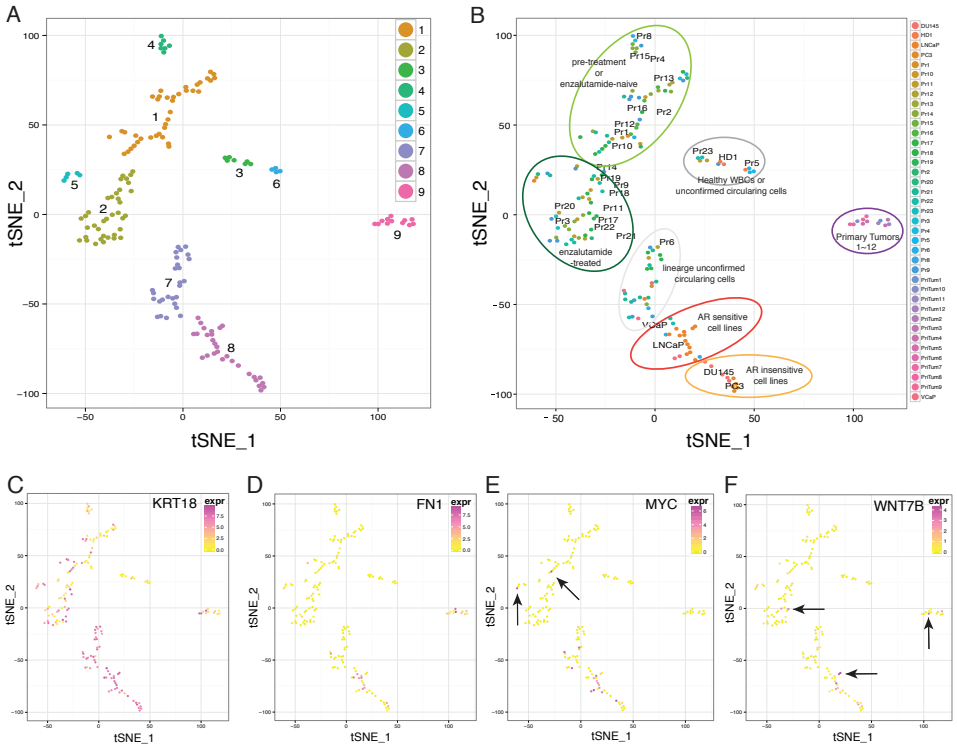

Figure 4: CTC dataset. A. Unsupervised classification of 169 cells including circulating tumor cells. Samples are clearly separated into four major groups: group-9, primary tumor; group-1 and group-2, cancer cell lines; group-3 and group-6, healthy control cells and group-8, cell lines. B. The color scheme in A is changed to reflect the cell origin, and the cell-to-cell heterogeniety is now apparant. Pr: patient CTC; PriTum: primary tumor; HD1: healthy white blood cells as control. C-F. Gene expression pattern for epithelial marker KRT18 (C), mesenchymal marker FN1 (D), oncogene MYC (E) and a non-canonical Wnt pathway gene WNT7B (F). The arrows point to rare cells that express MYC or WNT7B, such cell-to-cell heterogeniety might shed light on the mechanisms of treatment resistance.

4 Conclusion

We present Hastings, a versatile and user-friendly R pipeline for large-scale RNA-Seq data analysis. In all datasets tested here, Hastings identified correctly major groups and interestingly many sub-groups. Moreover, the high dimensional data is visualized in 2D and

potential marker genes are identified. Furthermore, the intra-sample heterogeneity can easily be inspected for each gene on top of the grouping structure. Hastings is not only easy to use, it can also be easily adopted and extended. For example, the data filtering step could be modified for special requirements from individual experimental design, and it might be also helpful to designate a pre-defined HVG set. We believe that Hastings could be widely applied to various large-scale studies for structure analysis and visualization.

5 Acknowledgement

References

[1] Katherine A Hoadley, Christina Yau, Denise M Wolf, Andrew D Cherniack, David Tamborero, Sam Ng, Max DM Leiserson, Beifang Niu, Michael D McLellan, Vladislav Uzunangelov, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, 158(4):929–944, 2014.

[2] Devon A Lawson, Nirav R Bhakta, Kai Kessenbrock, Karin D Prummel, Ying Yu, Ken Takai, Alicia Zhou, Henok Eyob, Sanjeev Balakrishnan, Chih-Yang Wang, et al. Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. *Nature*, 2015.

[3] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

[4] Laurens Van Der Maaten. Accelerating t-sne using tree-based algorithms. *The Journal of Machine Learning Research*, 15(1):3221–3245, 2014.

[5] Nathan Halko, Per-Gunnar Martinsson, Yoel Shkolnisky, and Mark Tygert. An algorithm for the principal component analysis of large data sets. *SIAM Journal on Scientific computing*, 33(5):2580–2594, 2011.

[6] Gad Abraham and Michael Inouye. Fast principal component analysis of large-scale genome-wide data. *PloS one*, 9(4):e93766, 2014.

[7] Ricardo JGB Campello, Davoud Moulavi, and Joerg Sander. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*, pages 160–172. Springer, 2013.

[8] Dmitry Usoskin, Alessandro Furlan, Saiful Islam, Hind Abdo, Peter Lönnerberg, Daohua Lou, Jens Hjerling-Leffler, Jesper Haeggström, Olga Kharchenko, Peter V Kharchenko, et al. Unbiased classification of sensory neuron types by large-scale single-cell rna sequencing. *Nature neuroscience*, 18(1):145–153, 2015.

[9] David T Miyamoto, Yu Zheng, Ben S Wittner, Richard J Lee, Huili Zhu, Katherine T Broderick, Rushil Desai, Douglas B Fox, Brian W Brannigan, Julie Trautwein, et al. Rna-seq of single prostate ctcs implicates noncanonical wnt signaling in antiandrogen resistance. *Science*, 349(6254):1351–1356, 2015.