Translational Bioinformatics 5
Series Editor: Xiangdong Wang, MD, PhD, Prof

Xiangdong Wang *Editor*

Single Cell Sequencing and Systems Immunology

Springer

Translational Bioinformatics

Volume 5

Series editor
Xiangdong Wang, MD, Ph.D.
Professor of Medicine, Zhongshan Hospital, Fudan University Medical School, China
Director of Shanghai Institute of Clinical Bioinformatics, (www.fuccb.org)
Professor of Clinical Bioinformatics, Lund University, Sweden

Aims and Scope

The Book Series in Translational Bioinformatics is a powerful and integrative resource for understanding and translating discoveries and advances of genomic, transcriptomic, proteomic and bioinformatic technologies into the study of human diseases. The Series represents leading global opinions on the translation of bioinformatics sciences into both the clinical setting and descriptions to medical informatics. It presents the critical evidence to further understand the molecular mechanisms underlying organ or cell dysfunctions in human diseases, the results of genomic, transcriptomic, proteomic and bioinformatic studies from human tissues dedicated to the discovery and validation of diagnostic and prognostic disease biomarkers, essential information on the identification and validation of novel drug targets and the application of tissue genomics, transcriptomics, proteomics and bioinformatics in drug efficacy and toxicity in clinical research.

The Book Series in Translational Bioinformatics focuses on outstanding articles/chapters presenting significant recent works in genomic, transcriptomic, proteomic and bioinformatic profiles related to human organ or cell dysfunctions and clinical findings. The Series includes bioinformatics-driven molecular and cellular disease mechanisms, the understanding of human diseases and the improvement of patient prognoses. Additionally, it provides practical and useful study insights into and protocols of design and methodology.

Series Description

Translational bioinformatics is defined as the development of storage-related, analytic, and interpretive methods to optimize the transformation of increasingly voluminous biomedical data, and genomic data in particular, into proactive, predictive, preventive, and participatory health. Translational bioinformatics includes research on the development of novel techniques for the integration of biological and clinical data and the evolution of clinical informatics methodology to encompass biological observations. The end product of translational bioinformatics is the newly found knowledge from these integrative efforts that can be disseminated to a variety of stakeholders including biomedical scientists, clinicians, and patients. Issues related to database management, administration, or policy will be coordinated through the clinical research informatics domain. Analytic, storage-related, and interpretive methods should be used to improve predictions, early diagnostics, severity monitoring, therapeutic effects, and the prognosis of human diseases.

Recently Published and Forthcoming Volumes

Applied Computational Genomics
Editor: Yin Yao Shugart
Volume 1

Pediatric Biomedical Informatics
Editor: John Hutton
Volume 2

Bioinformatics of Human Proteomics
Editor: Xiangdong Wang
Volume 3

Bioinformatics for Diagnosis, Prognosis and Treatment of Complex Diseases
Editor: Bairong Shen
Volume 4

More information about this series at http://www.springer.com/series/11057

Xiangdong Wang
Editor

Single Cell Sequencing and Systems Immunology

Honor editors
Xiaoming Chen
Zhihong Sun
Jinglin Xia

Springer

Editor
Xiangdong Wang
Zhongshan Hospital
Fudan University
China

Shanghai Institute of Clinical Bioinformatics
China

ISSN 2213-2775 ISSN 2213-2783 (electronic)
Translational Bioinformatics
ISBN 978-94-017-9752-8 ISBN 978-94-017-9753-5 (eBook)
DOI 10.1007/978-94-017-9753-5

Library of Congress Control Number: 2015936190

Springer Dordrecht Heidelberg New York London

Printed on acid-free paper

Springer Science+Business Media B.V. Dordrecht is part of Springer Science+Business Media (www.springer.com)

Chapter 2
Uncovering Phenotypes with Supercells: Applications to Single-Cell Sequencing

Julián Candia, Jayanth R. Banavar, and Wolfgang Losert

Abstract The so-called "Supercell paradigm" is a method for phenotyping based on single-cell multidimensional data, which has been recently proposed by the authors of this Chapter and collaborators within the larger context of single-cell biology. Supercells are multidimensional objects that represent the collective behavior of groups of cells and carry a distinct phenotype, which is often obscured at the single-cell level due to high cell-to-cell variability. The Supercell framework provides a quantitative assessment of the critical sample size and the number of simultaneous single-cell measurements needed to build a phenotype, which is a key piece of information given the fact that, in many single-cell applications, the number of measured cells and the number of measurements per cell are severely limited due to a variety of constraints, such as experimental costs, technological capabilities, specimen collection procedures, the availability of specialized personnel, and others. In this Chapter, we review the Supercell method and explore the potential for its application to single-cell sequencing datasets.

Keywords Single-cell biology • Single-cell genomics • Cell heterogeneity • Machine learning • Supercells

2.1 Introduction

Heterogeneity from cell to cell is now widely recognized as a key feature of many living systems, which enables their adaptation to changing environmental conditions (Altschuler and Wu 2010). Moreover, similar mechanisms appear to play a significant role in a tumor's ability to survive, proliferate, spread and resist therapy (Marte 2013). Single-cell heterogeneity is often encountered in biomedical research, as well as in the clinical realm, and leads to particular challenges for studies that

J. Candia, Ph.D. (✉)
Center for Human Immunology, National Institutes of Health, 10#7N115,
9000 Rockville Pike, Bethesda, MD 20892, USA
e-mail: julian.candia@nih.gov

J.R. Banavar, Ph.D. • W. Losert, Ph.D.
Department of Physics, University of Maryland, College Park, MD 20742, USA

X. Wang (ed.), *Single Cell Sequencing and Systems Immunology*,
Translational Bioinformatics 5, DOI 10.1007/978-94-017-9753-5_2

are based on a limited number of single cells. One important example is provided by state-of-the-art single-cell genomics technologies, which enable measuring the expression level of all genes in a single cell. However, the number of cells for which all genes can be measured is limited by both cost and instrument capacity. For these new high-dimensional data with limited numbers of data points, data analysis methods that rely on high-dimensional clustering procedures, Gaussian mixture approximations, and other standard classification techniques may be expected to fail. Therefore, it is of paramount importance to address the problem of phenotypic classification when single cells are highly heterogeneous and the number of cells available is small.

Within this context, we have recently proposed the so-called 'Supercell Paradigm' (Candia et al. 2013, 2014) as a general method for single-cell phenotyping that focuses on emergent properties of groups of cells. The key contribution of this method is to provide a quantitative assessment of the critical sample size and number of simultaneous single-cell measurements needed to identify a phenotype with strong predictive power. In (Candia et al. 2013), the Supercell framework was developed and applied to datasets obtained by imaging of cell nuclei and multicolor flow cytometry, as illustrations of the potential of this method to be applied to build multi-parametric phenotypes from different single-cell technologies.

The purpose of this Chapter is to review the Supercell method in detail and to explore the potential for its application in the context of single-cell sequencing data. In order to motivate the need for novel methods of analysis, Sect. 2.2 briefly overviews the challenges arising from high-dimensional single-cell technologies. In Sect. 2.3, we introduce the Supercell paradigm and illustrate the rationale of the method with some examples. In Sect. 2.4, we show the application of the method to single-cell RNA-seq datasets and discuss the potential for further applications. Finally, our Conclusions are stated in Sect. 2.5.

2.2 Phenotypic Heterogeneity and Small-Sample Effects: The Single-Cell Challenge

During the progression from the zygotic stage to adulthood, the aggregate effects of numerous somatic mutations result in the occurrence of several cell lineages with different genotypes in one individual, a phenomenon described as mosaicism (Lupski 2013). Although the true extent of such mosaicism is yet unknown, this phenomenon appears ubiquitous and has led scientists to speculate that each cell in the human body may have a unique genomic signature (Lupski 2013; Shapiro et al. 2013). Many of these mutations are expected to be neutral and others may be disadvantageous and go extinct. Some of them may even be beneficial: For instance, the widespread somatic mutations in the brain, observed in the form of aneuploidy or retro-transposon insertions, might contribute to normal brain

function (Baillie et al. 2011; Evrony et al. 2012). However, other somatic mutations are instrumental for the physiologic process of aging (Lopez-Otin et al. 2013) and for the onset of cancer (Tomasetti et al. 2013) and other diseases.

Indeed, single-cell heterogeneity poses challenges as well as huge opportunities in the development and improvement of strategies for the diagnosis and treatment of many diseases (Speicher 2013). For instance, Beckman et al. (2012) have very recently assessed the impact of single-cell heterogeneity, as well as that of genetic instability, in the development of effective nonstandard strategies for personalized cancer treatment. Manifestations of cell heterogeneity in healthy and diseased cell samples have ubiquitously been reported in the growing field of single-cell biology, ranging from human pluripotent embryonic stem cell cultures (de Souza 2012; Tang 2012; Drukker et al. 2012) and apoptosis mechanisms in cancer cell lines (Schmid et al. 2012), to reversible adaptive plasticity in tumors such as human neuroblastoma (Chakrabarti et al. 2012) and pressure-driven shape features of *C. elegans* embryonic cells (Fujita and Onami 2012). For recent reviews of the impact of tumor heterogeneity at different levels (genetic, epigenetic, the tumor microenvironment, the immune response, and other factors such as diet and the microbiota), see Meacham and Morrison (2013), Burrell et al. (2013), Junttila and de Sauvage (2013), and Bedard et al. (2013).

Besides the inherent biological variability from cell to cell, an additional layer of heterogeneity arises from technical noise. Indeed, the accuracy and reliability of single-cell analysis is severely limited by whole-genome and whole-transcriptome amplification noise from a variety of sources (Macaulay and Voet 2014). Although further innovations will be needed to develop the capacity to directly sequence unamplified DNA and RNA derived from single cells, direct library preparation from single-cell genomes has been demonstrated (Falconer et al. 2012; Falconer and Lansdorp 2013) and direct sequencing of single molecules is already a possibility for DNA and RNA (Ozsolak et al. 2009; Coupland et al. 2012).

Another limiting characteristic of current single-cell sequencing studies is the small number of cells investigated, typically in the range from tens to a few hundreds. Within the broader realm of single-cell biology, the inability to have large samples often arises due to technical limitations and cost considerations as well as the nature of the biological/clinical problem at hand. In the field of stem cell research, for instance, stem cells are extremely rare. Thus, identifying and sorting stem cells through flow cytometry yields, even at best, only limited numbers of cells. As an example, long-term hematopoietic stem cells (LT-HSCs) identified via immunophenotypes such as $Lin^{-}Kit^{+}Sca^{+}CD34^{lo}Flt3^{-}$ (Christensen and Weissman 2001) and SLAM (Kiel et al. 2001) represent only about 0.0075 % of the cells from whole bone marrow specimens; thus, more than a million whole bone marrow cells need to be extracted, stained with multiple fluorochromes and sorted in order to yield about one hundred LT-HSCs.

In this Chapter, we will focus on the important case where the number of measured cells is limited for one of the reasons listed above. Furthermore, we will consider situations where cell behavior is so heterogeneous that distinct cell populations have overlapping distributions. Figures 2.1 and 2.2 show schematic

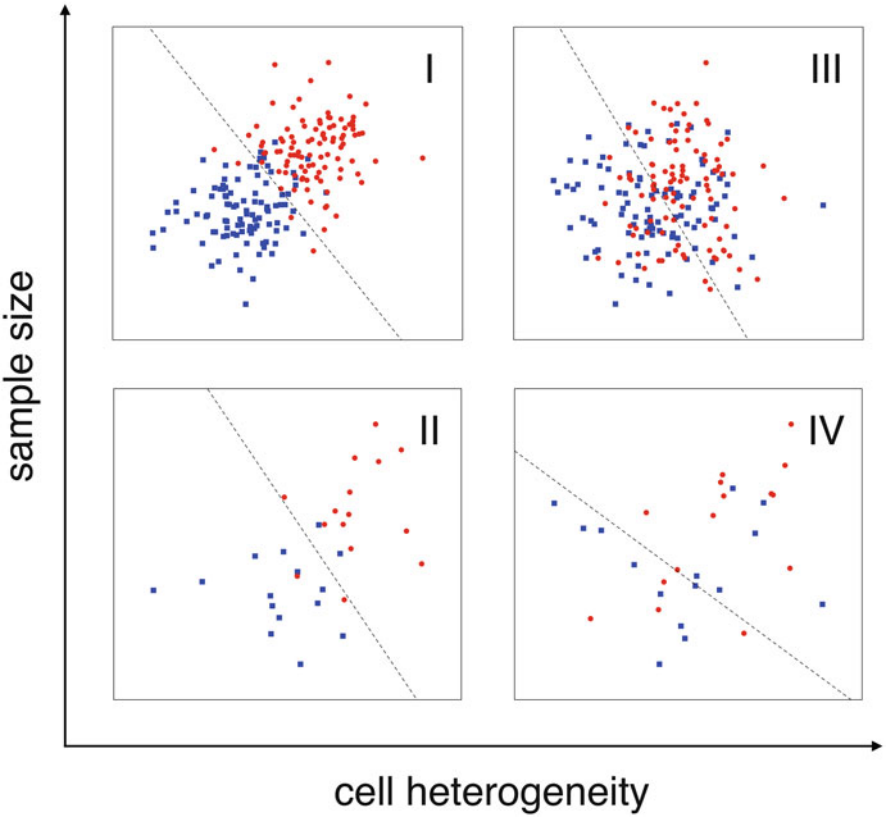

Fig. 2.1 Schematic representation of four scenarios that result from the combinations of low/high cell heterogeneity with small/large sample size. Within each scenario, a *dashed line* shows the linear boundary that optimally separates the two classes (represented by *blue squares* and *red circles*) using a Support Vector Machine (SVM)

representations of the relations between cell heterogeneity, sample size, and the expected classification accuracy of training and test observations. We consider different scenarios in which single-cell measurements are performed on cells that belong to one of two possible classes (i.e. distinct biological phenotypes, such as e.g. cells from a cancer cell line compared with cells from a healthy cell line). On the one hand, cell heterogeneity refers to the observed overlap between the two cell populations, which arise from the biology (depending on how well the chosen biomarkers can inherently distinguish one cell population from another), from the technical procedure (e.g. measurement noise, batch effects, etc.), or more generally from a combination of both biological and technical considerations. On the other hand, sample size refers to the number of single cells measured, which typically depends on considerations of cost, instrument capacity and the number of cells of a particular type. Figure 2.1 schematically considers four main scenarios that

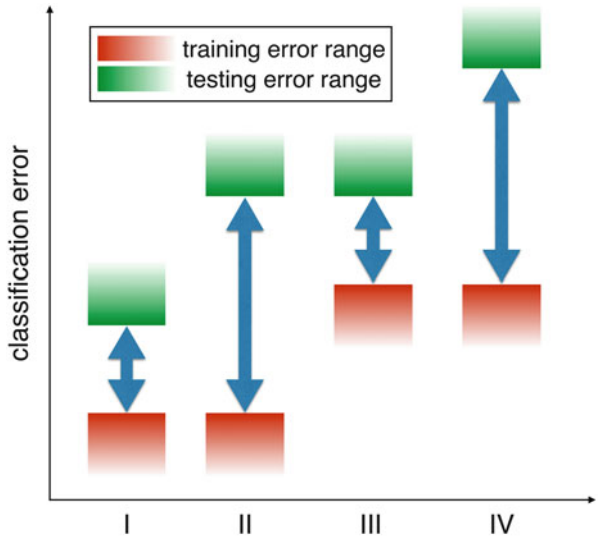

Fig. 2.2 Schematic representation of the expected training (learning) and testing (prediction) classification errors for each of the four scenarios shown in Fig. 2.1. The low-heterogeneity cases (I and II) lead to good class separation and thus small training errors, whereas the high-heterogeneity cases (III and IV) exhibit poorer class separation and larger training errors. On the other hand, large sample sizes (cases I and III) yield a small classification error increase in going from the training to the testing phase, while the increase in more significant with smaller sample sizes (cases II and IV)

result from the combinations of low/high cell heterogeneity with small/large sample size. Within each scenario, a dashed line shows the linear boundary that optimally separates the two classes using a Support Vector Machine (SVM), one popular machine learning method that is widely used as a supervised classifier (more details on SVMs will be given in Sect. 2.3). In supervised classification, the classifier is first built during a so-called *training* or *learning phase*, in which we must know in advance the true classification for each cell. The ability for the classifier to correctly separate the measurement hyperspace in two regions that reflect the true separation between cell classes is quantified by means of the learning error, i.e. the percentage of cells that lie in the wrong side of the classification boundary. Naturally, the learning error is smaller when the cell heterogeneity is smaller, as in Cases I and II, since then it's possible to draw a linear boundary that separates very well the two classes and misclassifies just a few cells. Ranges of expected training errors are schematically depicted in Fig. 2.2 by red-colored fading regions: they are expected to be low in Cases I and II, and higher in Cases III and IV, which are correspondingly characterized by larger cell heterogeneity.

Supervised classifiers are often intended as methods to predict the correct class of new (unknown) instances. This can be used to determine the primary site of a metastatic cancer or to diagnose a disease in a patient from cells obtained with non-invasive or minimally invasive procedures, among many possible applications.

In these examples, it is vital to assess the ability of the method to correctly predict the cell class in the so-called *prediction* or *testing phase*. But, even if we are just interested in inference (that is, to learn the relations between biomarkers to describe phenotypes, rather than to make predictions based on them), it is important to evaluate the classifier's performance during the testing phase to assess whether the classifier has adequately captured the true patterns of classification. Indeed, it is possible to have a classifier that performs well on the training data because it follows the training class labels very closely, but then fails in predicting new instances. This phenomenon is called *overfitting*.

As shown in Fig. 2.2, the expected testing errors given by the green-colored fading regions are always, on average, larger than the corresponding learning errors. However, the increase from learning to testing errors depends on how representative the learning samples are of the true class distributions, which naturally depends on sample size. Thus, we expect large sample sizes (Cases I and III) to yield small training-to-testing error increases, whereas for small sample sizes (Cases II and IV) we expect much larger training-to-testing error increases.

It is interesting to note that, whereas Case I is clearly the best scenario and Case IV is the worst, Cases II and III may yield comparable performance. Yet, there is a delicate balancing act to negotiate the trade-off between effective sample size and effective class heterogeneity in order to find the optimal sweet spot between the two, which yields the optimal performance for a given experimental single-cell dataset. In the next Sections, we develop these ideas further and apply them to both synthetic and true single-cell genomics data.

2.3 The Supercell Paradigm

Highly heterogeneous cell populations (as those represented by Cases III and IV in Fig. 2.1) are not linearly separable and, therefore, the boundary that separates them is ill-defined. In this situation, one solution is to adopt curved decision boundaries that may better fit the separation between classes. To this end, a variety of machine-learning methods such as support vector machines (SVMs) with non-linear kernels, K-nearest neighbors, quadratic and higher-order discriminant analysis, and others, are available to find non-linear class boundaries in high-dimensional measurement space (Hastie et al. 2009; Garteh et al. 2013). However, those methods are also prone to overfitting, a well-known phenomenon in which the classifier performs well on the training data merely because it was trained to follow closely the training instances, but then fails in predicting new instances. As an alternative approach, we build supercell samples as mathematical objects that can be treated in the same way as the directly measured single-cell samples, but which use well-known properties of statistical ensembles to enhance the separation between cell subpopulations. Then, we apply low-variance machine learning methods (such as, e.g., SVM with a linear kernel) on those supercell samples.

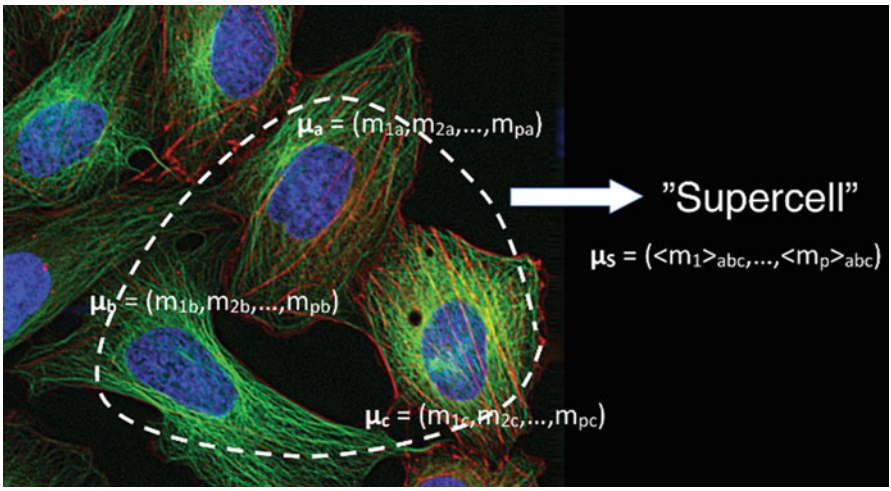

Fig. 2.3 Schematic representation of the supercell averaging procedure. On each cell, p parameters are measured; each cell is represented by a measurement vector $\boldsymbol{\mu}$. A supercell of size N is calculated by randomly selecting N single cells from the sample and then averaging their measurement vectors. By repeating this procedure, we obtain a sample of supercells from the original sample of measured single cells

In order to capture multidimensional cell phenotypes, a 'supercell of size N' is defined as the average of the individual measurement vectors of a group of N randomly chosen cells. By repeatedly taking different random subsets of N cells, 'supercell samples' can be built out of the original single-cell datasets. This procedure is schematically represented in Fig. 2.3, where p parameters are measured on each cell, which is thus quantitatively characterized by a measurement vector $\boldsymbol{\mu}$ with p components. In its simplest realization, supercell averaging proceeds by taking N cells at random and averaging their measurement vectors into a supercell vector $\boldsymbol{\mu}_S$. Since the single-cell sample size, N_s, is usually small, supercell averaging is typically performed by selecting cells at random *with replacement*, that is, allowing the same single cell to be chosen more than once. This procedure is indeed similar to the well-known method of *bootstrapping* (Garteh et al. 2013). By iterating this procedure, we obtain a representative sample of N_s' supercells out of the original sample of N_s single cells. Notice that, although the simplest approach builds supercells by combining single cells chosen at random, it is also possible to incorporate additional information to the cell averaging process. In the case of high content multiplexed tissue imaging, for instance, the available 2D or 3D spatial information (localization of each cell, orientation relative to its neighbors, the microenvironment and surrounding extra-cellular matrix, etc.) could be used as inter-cellular level information in the generation of supercells. Analogously, cell cycle phase, cell subtype, etc. may be incorporated to the supercell averaging process for datasets generated by other single-cell technologies such as single-cell genomics.

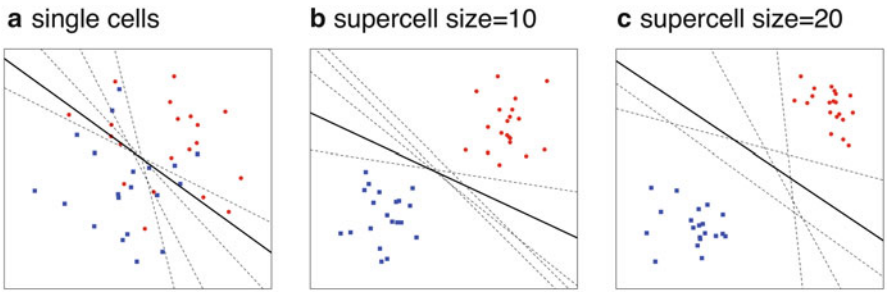

Fig. 2.4 Class separation of 2D synthetic datasets. Samples of 20 cells were obtained for each class, which were randomly generated from uncorrelated 2D normal distributions. The *thick solid line* shows the linear SVM class boundary for the samples displayed (*blue squares* and *red circles*). By generating new samples (not shown), different boundaries are obtained, which are displayed as *thin dashed lines*. (**a**) Learning with single cells, the two populations are not linearly separable. (**b**) Linear separation is achieved by using supercells of size 10. (**c**) By increasing the supercell size to 20, the class separation increases but becomes less robust due to overfitting

After cell averaging, machine learning is used to learn what combination of parameters best distinguishes the different phenotypes. The method implemented in Candia et al. (2013) is a support vector machine (SVM) with a linear kernel, but it can be extended to non-linear mappings that may better reflect the inherent structure of the data. In the linear case, the components of the vector normal to the boundary hyperplane can be straightforwardly interpreted as amplitudes that determine the relative significance of the measured parameters in achieving class separation. Moreover, by introducing appropriate quality functions to balance the tradeoff between separation and robustness, the Supercell paradigm is able to assess the optimal supercell size in order to achieve phenotypic classification when single cells are highly heterogeneous and the number of cells available is small.

Figure 2.4 shows an illustration of the supercell method on 2D synthetic datasets, in which samples of $N_s = 20$ cells were obtained for each class, which were randomly generated from uncorrelated 2D normal distributions. The distributions have the same shape and variance, but their centers are separated. When considering these distributions at the single-cell level, we see that the samples are highly overlapping with no well-defined class boundary (Fig. 2.4a). The thick solid line shows the best class separation obtained from a linear SVM applied to the samples displayed in the figure, where the two classes are represented by blue squares and red circles, respectively. In order to represent the fluctuations arising from the combination of cell population overlap and small sample size, the dashed lines show SVM class boundaries obtained by generating new samples of 20 cells each (these additional samples are not displayed in the figure, only the resulting class boundaries are). By generating supercell samples, the cell populations separate, as expected on the basis of the central limit theorem. For simplicity, we choose a supercell sample size, N_s', equal to the original single-cell sample size $N_s = 20$. Figure 2.4b shows the linear separation achieved by using supercell size $N = 10$. As before, fluctuations arising from different samples are shown by dashed lines. In Fig. 2.4c, the supercell

size has been increased to $N = 20$. Correspondingly, the class separation is even larger, at the expense of a larger variation in the orientation of the class boundaries as different cell samples are considered. This is a manifestation of overfitting, since in this case the supercells used to learn the class boundary are highly correlated with each other and very close to the overall class means.

It should be noted that the orientation of the class boundary is affected by different factors. One of them is single-cell sampling: from each class distribution, only N_s single cells are actually measured. When N_s is relatively small, strong sampling effects are to be expected and thus, large boundary fluctuations will arise from considering different sets of single-cell ensembles. On the other hand, when generating supercell samples out of the measured single cells, there are many different ways of choosing N single cells to average together into each supercell, which leads to an additional source for class boundary fluctuations. Finally, it is important to notice that the supercell averaging procedure has the combined effect of shrinking the cell distribution in parameter space, as well as that of modifying the shape of the distribution closer to a normal distribution, which is the expected effect of averaging due to the central limit theorem. Thus, if the original single-cell distribution is skewed, fat-tailed, or has many outlier observations beyond the expected range of a normal distribution, the shape of the resulting supercell distribution will be significantly different from the corresponding single-cell one.

The orientation of the class boundary conveys important information about the relative importance of the different measures. Indeed, when the number of measured parameters per cell, p, is very large, the orientation of the class boundary allows us to rank-order the measures, remove the least significant one, and reiterate the learning procedure, a process known as *recursive feature elimination* (Guyon et al. 2006). Thus, based on the considerations mentioned above, it is important to stress the fact that supercell averaging allows us to optimally characterize cell phenotypes based on class labeling and, thus, to work around the difficulties imposed by cell heterogeneity within each class.

These phenotypes are collective properties of cells within each class and do not necessarily reflect the best combination of parameters to characterize single cells within each class.

In Candia et al. (2013), we have developed and applied the Supercell/SVM paradigm to datasets obtained by different single-cell technologies, e.g. imaging of cell nuclei and multicolor flow cytometry. As a case example of the latter, we focused on the challenging problem of building molecular phenotypes to characterize the differences between two non-infectious uveitides (the ocular manifestations of sarcoidosis and Behçet's disease), which are very difficult to diagnose in the clinic and require different treatments. By performing two scattering and 14 fluorescent measurements on each cell, samples from 7 sarcoidosis and 6 Behçet's patients were measured. Since the cohort was small, prediction testing was carried out by a jackknife (leave-one-out) cross-validation procedure. The SVM boundary allows one to rank-order the 16 measures from most to least significant, according to the components of the vector normal to the hyperplane that separates the two diseases. Thus, one can selectively remove the least significant measurements from the list

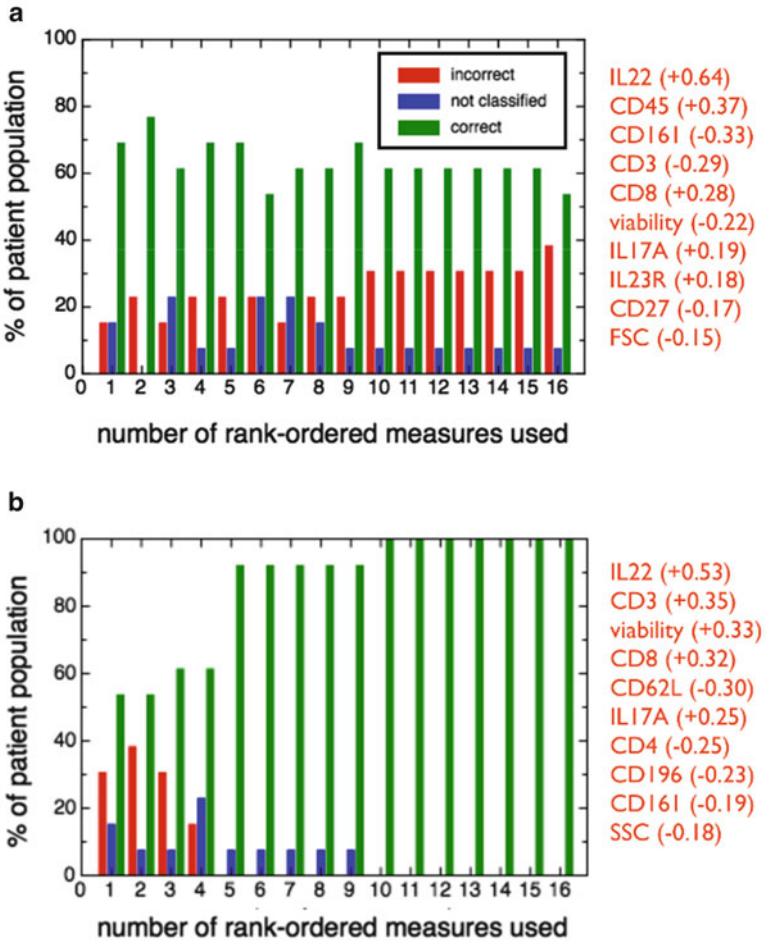

Fig. 2.5 Example of supercell phenotyping using multicolor flow cytometry. Leave-one-out (jackknife) cross-validation results for sarcoidosis versus Behçet's disease using supercells of size N = 500, where each patient is represented by a cloud of 100 supercells, as a function of the number of rank-ordered measures used: (**a**) All cells; (**b**) CD8+ T cells. The *bars* show percentages of correct (*green*), unclassified (*blue*) and incorrect (*red*) predictions. To the right of each panel, the list of the top 10 rank-ordered measures is shown (Adapted from Candia et al. 2013)

and explore the minimal number of measures needed to correctly predict the class of all (or at least most of) the samples. Figure 2.5 shows jackknife results for supercells of size N = 500, where each patient is represented by a cloud of 100 supercells, as a function of the number of rank-ordered measures used. The list of the top 10 measures is shown to the right of each panel. Figure 2.5a shows jackknife results based on all cells, while Fig. 2.5b displays results based on $CD8^+$ T cells, a subpopulation that can be determined by manual gating ($CD3^+$ $viab^-$ $CD8^+$ $CD4^-$) and typically represents about 5 % of the peripheral blood sample.

Since each patient is represented by a cloud of supercells, a prediction was made only when more than 95 % of those supercells lie on any one side of the SVM boundary. Correct predictions are shown by green bars, incorrect predictions by red bars, while unclassified samples are shown in blue. While predictions based on all cells are very poor, for $CD8^+$ T cells no failed predictions are incurred when five or more measures are used. Therefore, the top five measures listed in Fig. 2.5b, can be linearly combined in order to be used on $CD8^+$ T cells as molecular phenotypes that distinguish the two diseases.

2.4 Applications of the Supercell Framework to Single-Cell Sequencing: A Case Study

In this Section, we will work out a case study using publicly available single-cell RNA-seq datasets. The purpose of this Section is to illustrate possible applications of the ideas discussed earlier to the kinds of datasets produced by state-of-the-art single-cell sequencing technologies. Our main focus here is not on the biology, but rather, on the method and its potential as an analysis tool on datasets characterized by highly-overlapping phenotypes obtained from highly-dimensional single-cell datasets that span a limited number of cells, typically around or below 100 cells.

To this end, we will use publicly available data (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE52583) that correspond to microfluidic single-cell RNA-seq on 198 individual mouse lung epithelial cells at four different stages through development, namely E14.5, E16.5, E18.5, and AT2 (adult). In order to control background and normalization, these datasets include 92 external RNA (ERCC) spike-ins; moreover, one no-cell and two 200-cell bulk control samples were generated for time point E18.5 (see http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE52583; Treutlein et al. 2014 for further technical details). In our analysis, we used 45 E14.5 cells, 27 E16.5 cells, 34 E18.5 cells (which correspond to replicate # 2 in http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE52583) and 46 AT2 (adult) cells, totaling 152 cells. Based on the datasets in (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE52583), Treutlein et al. (2014) have very recently confirmed the basic outlines of the conventional model of cell type diversity in the distal lung, as well as discovering a large number of novel transcriptional regulators and cell type markers that discriminate between different cell populations.

After obtaining RNA-seq expression values in terms of Fragments Per Kilobase of transcript per Million mapped reads (FPKM), we transformed them to $\log_2(\text{FPKM} + 0.5)$. The RNA-seq matrix is typically very sparse, with most entries corresponding to zero transcripts (which, in our log-transformed scale, are represented by -1 values). Thus, single-cell genomics poses particular challenges for data analysis due to low signal-to-noise ratios, small sample sizes, and with the additional complication arising from their attributes spanning an extremely high-dimensional space.

The issue of high-dimensional spaces can be addressed by implementing one of several possible feature selection schemes (Guyon et al. 2006). In Sect. 2.3, we showed one example in which a panel of multiple flow cytometry measures was first used to achieve the separation of phenotypes, then followed by *recursive feature elimination* (also called *backward stepwise selection*) to sequentially remove the least important markers in a top-down fashion until a core set of markers was found. Alternatively, one could start by considering individual measures and selecting the one that has the smallest classification error, and iterating in a forward, stepwise manner, adding more predictors to the model, one at a time. This bottom-up procedure is called *recursive feature addition* or *forward stepwise selection*. Other approaches to feature selection are hybrid implementations that add predictors sequentially, in analogy to forward selection, but at each step may also remove measures that no longer provide an improvement in the model classification.

In this Section, however, we use a different approach. We focus on biologically meaningful gene sets that rely on relevant pathway information. This approach is complementary to fully unbiased methods of gene set selection that are agnostic to the biology. Indeed, whereas the former approach may lead to more focused knowledge in the context of specific biological processes, the latter may lead to the discovery of new molecular mechanisms and thus open up new avenues of research. Recalling that the RNA-seq data we are concerned with here correspond to individual mouse lung epithelial cells at different developmental stages, we will focus on a set of genes that have roles as oncogenes or tumor suppressor genes in KEGG pathways associated with small cell lung cancer (http://www.kegg.jp/kegg-bin/show_pathway?mmu05222) and non-small cell lung cancer (http://www.kegg.jp/kegg-bin/show_pathway?mmu05223). After disregarding two genes (Alk and Cdkn2a) that are uniformly undetected in all the 152 single-cell samples considered here, we are left with a panel of nine target genes, namely: Eml4, Fhit, Kras, Myc, Pten, Rarb, Rassf1, Rb1, and Trp53.

The SVM approach used in the previous Section dealt with a two-class learning problem based on multidimensional single-cell measurements in a nearly continuous range. The example discussed in more detail (Fig. 2.5) was based on multicolor flow cytometry intensity measurements, which typically lie above detection thresholds. In contrast, RNA-seq data are characterized by sparse expression matrices with many zeros and the learning problem we are considering here has four classes with a natural progression given by cell developmental stage. Rather than using SVM boundaries, this kind of classification problem is better solved by random forests, a generalization of decision trees in which instances are classified depending on a sequence of binary decisions based on measurement thresholds (Garteh et al. 2013; Breiman 2001). A variety of random forest algorithms has been successfully applied to many applications in bioinformatics (see e.g. Strobl et al. 2007 and references therein). Random forests can be applied to a wide range of prediction problems, even if they are nonlinear and involve complex high-order interaction effects, and they produce variable importance measures for each predictor variable.

Figure 2.6 shows results of random forests applied to the single-cell datasets described above. Based on so-called *out-of-bag* (OOB) data (see Garteh et al.

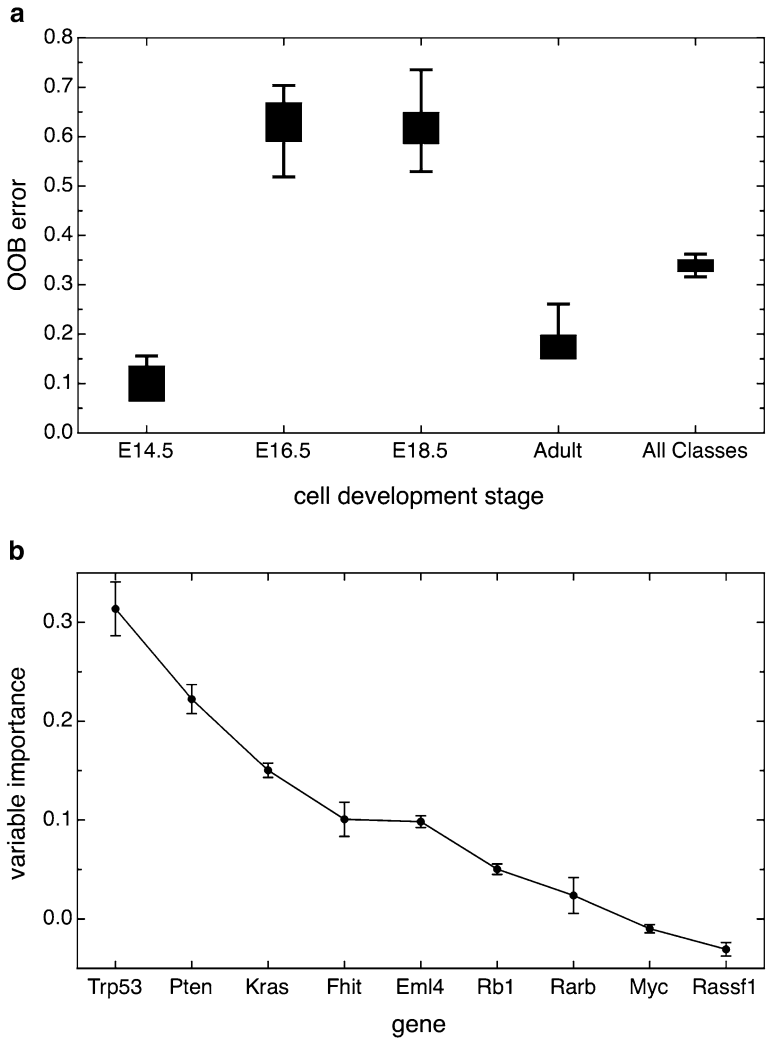

Fig. 2.6 Random forest results for RNA-seq individual mouse lung epithelial cells in 4 different stages of development: E14.5, E16.5, E18.5 and AT2 (adult). We focus on a set of 9 genes involved in lung cancer pathways. (**a**) *Box plots* showing so-called out-of-bag (OOB) error rate distributions (calculated from over 100 iterations of random forests built using 1,000 trees each) for each class. The overall mean OOB error rate for single-cell classification is 34 %. (**b**) Variable importance of each of the genes in the gene panel computed as the mean decrease of accuracy. Variable importance values have been normalized to add up to 1. Here, larger values indicate increased importance for the classification decision. The span of the vertical bars represents one standard deviation above and below the mean variable importance

2013; Breiman 2001 for background information and http://cran.r-project.org/web/packages/randomForest/index.htmlfor method implementation details), Fig. 2.6a shows the OOB error rates for the classification of cells in each of the four classes based on RNA-seq expression of genes in the selected gene panel. The box plot displays the distribution of OOB error rates over 100 iterations of random forests built using 1,000 trees each. We observe that, whereas the phenotypes for the early development stage E14.5 and the adult one appear well characterized (with mean OOB error rates of 11 % and 19 %, respectively), the intermediate E16.5 and E18.5 stages are poorly characterized (with mean error rates above 60 %). The overall mean OOB error rate for single-cell classification is 34 %. By recording the mean decrease of accuracy of predictions in the out-of-bag samples when a given predictor (gene) is excluded, we obtain a measure of the so-called *variable importance* for that predictor. Figure 2.6b shows the variable importance of each of the genes in the gene panel for random forest learning based on single cells. Variable importance values have been normalized to add up to 1. Notice that larger values indicate increased importance for the classification decision.

Now we can incorporate supercell averaging: following the rationale described in Sect. 2.3, we can generate a supercell ensemble of 45 E14.5 supercells, 27 E16.5 supercells, 34 E18.5 supercells and 46 AT2 (adult) supercells. As before, one supercell of size N is obtained by averaging the single-cell measurement vectors (in this case, associated with the expression of multiple genes) over N randomly chosen single cells with replacement (i.e. allowing the same cell to be chosen more than once). Based on one such supercell ensemble, we apply the random forest learning method using 1,000 trees. Then, we iterate this procedure 100 times and measure OOB error rate distributions, as we did on the (original) single cell datasets.

Figure 2.7a shows the mean OOB error resulting from random forests as a function of supercell size. As expected due to supercell averaging shrinkage via the central limit theorem, the mean OOB error decreases monotonically with supercell size. In order to choose the optimal supercell size, we need to implement a criterion to choose the appropriate degree of flexibility of our model. In other words, we need to optimize the so-called *bias-variance tradeoff*: as we average using supercells, the distributions shrink and their overlap decreases, making it easier to identify different classes; however, this decreased bias comes at the expense of an increased variance due to the introduction of effective correlations between supercells (i.e. supercell learning instances are not truly statistically-independent observations, as single cells are, and the equivalent sample size of effectively independent supercell observations becomes smaller than the original single-cell sample size). As a simple approach to adjust the OOB error rate to account for the model size (i.e. the choice of supercell size), we roughly estimate the prediction rate using an ad-hoc Information Criterion of the form $\mathrm{IC} = \mathrm{OOB\ error} + \mathrm{d}\ \sigma\ \sqrt{\mathrm{N}}$, where the second term is a penalty from using a high-dimensional parameter space of dimension d and supercells of size N. The $\sqrt{\mathrm{N}}$ dependence stems from the fact that, due to the central limit theorem, the width of supercell distributions shrinks as $\sqrt{\mathrm{N}}$, while σ represents an estimate of the overall variance of the error ε associated with each response measurement. In this case, we adopt $\sigma = 0.01$, while the dimensionality is $\mathrm{d} = 9$. Figure 2.7b shows IC as

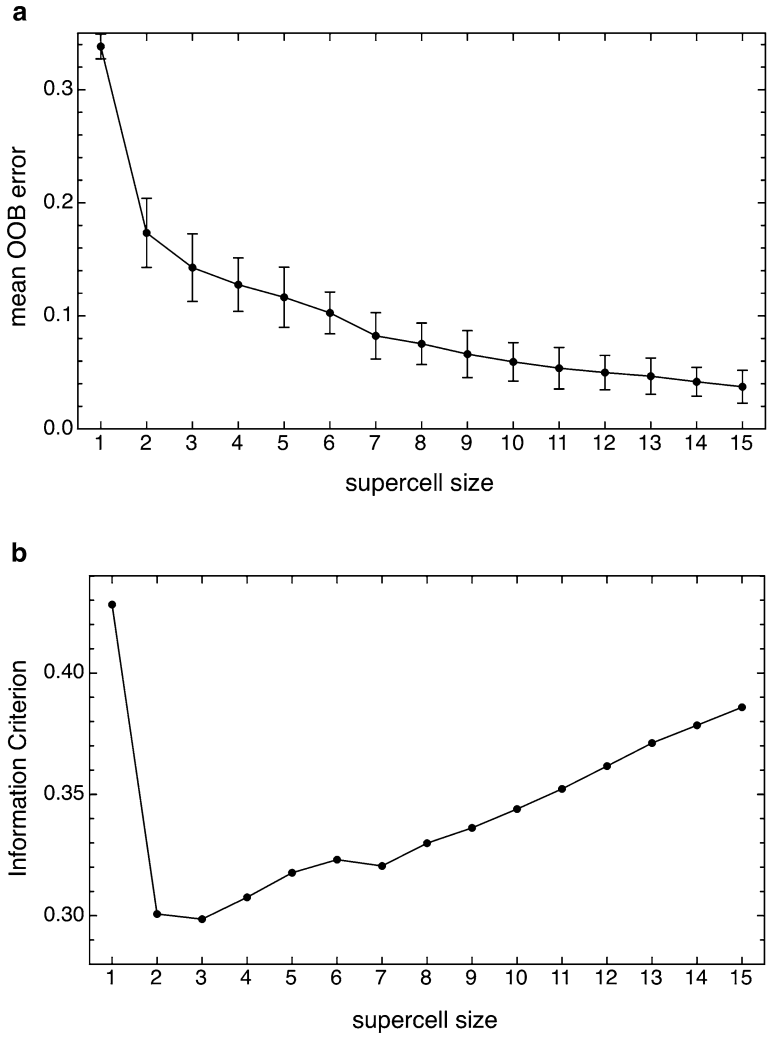

Fig. 2.7 Random forest results as a function of supercell size, using 1,000 trees in each random forest and averaging over 100 supercell realizations. (**a**) The mean OOB error decreases monotonically with supercell size, as expected due to supercell averaging shrinkage via the central limit theorem. (**b**) By considering an ad-hoc Information Criterion to balance the bias-variance tradeoff, the optimal supercell size is $N = 3$ (See text for details)

a function of supercell size. After adjusting by IC, we observe that the estimate for the prediction error is a minimum for $N = 3$; for $N > 3$, the penalty due to increased cell averaging overrides the gains obtained due to smaller OOB error rates, thus leading to increasing IC values. It should be pointed out that, with larger datasets, the validation set and cross-validation methods may be implemented to directly estimate prediction errors.

Finally, Fig. 2.8 shows random forest results for supercells of size $N = 3$. Figure 2.8a shows OOB error rate distributions (over 100 iterations of random forests built using 1,000 trees each) for each class. By comparing them to the single-cell OOB error rate distributions in Fig. 2.6a, we observe that, at the expense of small increases in the error rates of classes E14.5 and AT2, vast improvements in the classification of the intermediate development stages (E16.5 and E18.5) are achieved. The overall mean OOB error rate is also very significantly reduced. Figure 2.8b shows the importance of each of the genes in the gene panel computed as the mean decrease of accuracy. As before, variable importance values have been normalized to add up to 1 and larger values indicate increased importance for the classification decision. Standard deviations (shown by vertical bars) are larger in the supercell case compared with the single-cell case, as expected from the fact that, in the former, we average over different decision trees in the random forest as well as over different supercell realizations. The assessed relative importance of the various genes in the panel, however, does not display any significant differences.

2.5 Conclusions

The Supercell paradigm is a method for phenotyping based on single-cell multidimensional data, which has been recently proposed by the authors of this Chapter and collaborators within the larger context of single-cell biology, focusing on applications to multicolor flow cytometry and high-content image-based phenotyping (Candia et al. 2013). Supercells are multidimensional objects that represent the collective behavior of groups of cells; within this approach, supercells represent the building blocks of healthy and diseased phenotypes. From a conceptual standpoint, this approach naturally incorporates emergent behavior and thus cell heterogeneity, usually regarded as a roadblock in the pursuit of characterizing single-cell-level behavior, becomes the fundamental conceptual unit to identify collective phenotypes. From a practical perspective, the Supercell framework provides a quantitative assessment of the critical sample size and the number of simultaneous single-cell measurements needed to build a phenotype, which is a key piece of information given the fact that, in many single-cell applications, the number of measured cells and the number of measurements per cell are severely limited due to a variety of constraints, such as experimental costs, technological capabilities, specimen collection procedures, the availability of specialized personnel, and others.

Single-cell sequencing technologies generate datasets that pose particular challenges for data analysis due to low signal-to-noise ratios, small sample sizes, and extremely high-dimensional predictor spaces. In this Chapter, we discussed ways in which supercells could provide useful conceptual and computational means to deal with some of those challenges. Hopefully, these tools and ideas will stimulate further work and will contribute to advance the emerging and very promising field of single-cell biology.

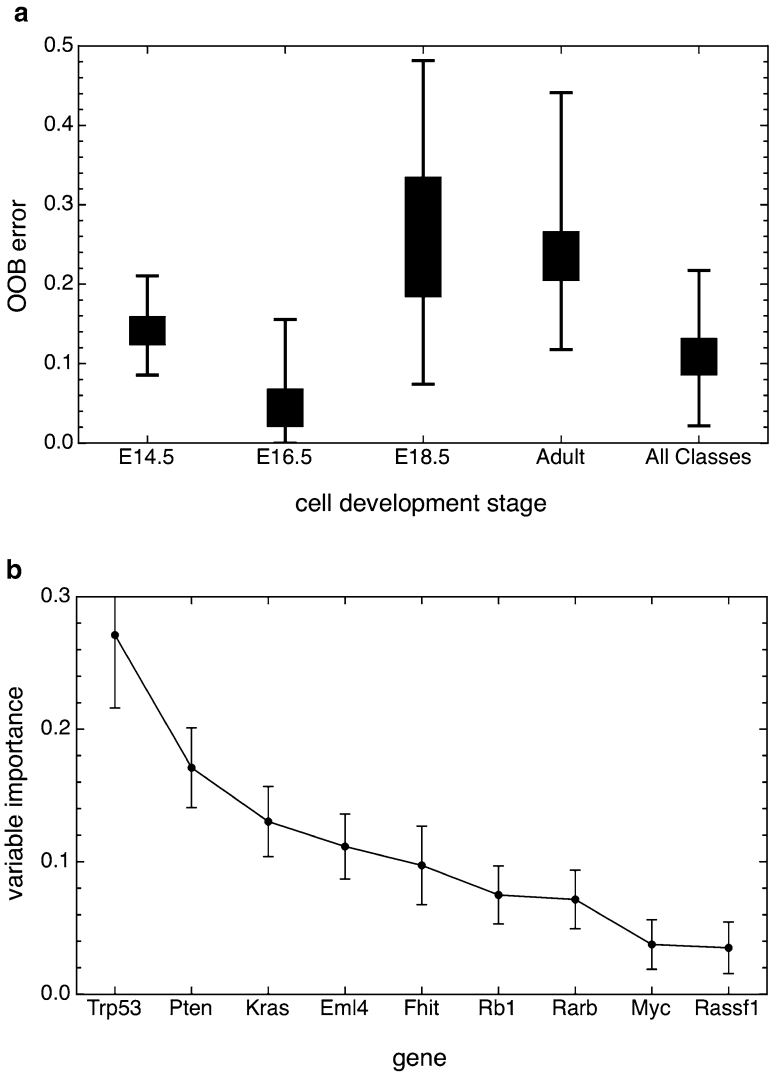

Fig. 2.8 Random forest results for supercells of size N = 3. (**a**) *Box plots* showing OOB error rate distributions (over 100 iterations of random forests built using 1,000 trees each) for each class. The overall mean OOB error rate is 14 %. (**b**) Variable importance of each of the genes in the gene panel computed as the mean decrease of accuracy. Variable importance values have been normalized to add up to 1. Here, larger values indicate increased importance for the classification decision. The span of the vertical bars represents one standard deviation above and below the mean variable importance

Acknowledgments We acknowledge our coauthors A. Biancotto, K. Cao, P. Dagur, M. Driscoll, A. Maritan, R. Maunu, J. P. McCoy Jr., R. B. Nussenblatt, H. N. Sen, and L. Wei, whose contributions to the Supercell approach (Candia et al. 2013) are extensively described in this Chapter. J. C. was supported by NIH Award Number T32CA154274 from the National Cancer Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health.

References

Altschuler S, Wu LF. Cellular heterogeneity: do differences make a difference? Cell. 2010;141:559.

Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, et al. Somatic retrotransposition alters the genetic landscape of the human brain. Nature. 2011;479:534–7.

Beckman RA, Schemmann GS, Yeang C-H. Impact of genetic dynamics and single-cell heterogeneity on development of nonstandard personalized medicine strategies for cancer. Proc Natl Acad Sci U S A. 2012;109(36):14586–91.

Bedard PL, Hansen AR, Ratain MJ, Siu LL. Tumour heterogeneity in the clinic. Nature. 2013;501(7467):355–64.

Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.

Burrell RA, McGranahan N, Bartek J, Swanton C. The causes and consequences of genetic heterogeneity in cancer evolution. Nature. 2013;501(7467):338–45.

Candia J, Maunu R, Driscoll M, Biancotto A, Dagur P, McCoy Jr JP, Sen HN, Wei L, Maritan A, Cao K, Nussenblatt RB, Banavar JR, Losert W. From cellular characteristics to disease diagnosis: uncovering phenotypes with supercells. PLoS Comput Biol. 2013;9:e1003215.

Candia J, Banavar JR, Losert W. Understanding health and disease with multidimensional single-cell methods. J Phys Condens Matter. 2014;26:073102.

Chakrabarti L, Abou-Antoun T, Vukmanovic S, Sandler AD. Front Oncol. 2012;2:82.

Christensen JL, Weissman IL. Flk-2 is a marker in hematopoietic stem cell differentiation: a simple method to isolate long-term stem cells. Proc Natl Acad Sci U S A. 2001;98(25):14541–6.

Coupland P, Chandra T, Quail M, Reik W, Swerdlow H. Direct sequencing of small genomes on the Pacific Biosciences RS without library preparation. Biotechniques. 2012;53:365–72.

de Souza N. Taming stem cell heterogeneity. Nat Method. 2012;9(7):645.

Drukker M, Tang C, Ardehali R, Rinkevich Y, Seita J, Lee AS, Mosley AR, Weissman IL, Soen Y. Isolation of primitive endoderm, mesoderm, vascular endothelial and trophoblast progenitors from human pluripotent stem cells. Nat Biotechnol. 2012;30(6):531–42.

Evrony GD, Cai X, Lee E, Hills LB, Elhosary PC, et al. Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. Cell. 2012;151:483–96.

Falconer E, Lansdorp PM. Strand-seq: a unifying tool for studies of chromosome segregation. Semin Cell Dev Biol. 2013;24:643–52.

Falconer E, Hills M, Naumann U, Poon SS, Chavez EA, et al. DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. Nat Methods. 2012;9:1107–12.

Fujita M, Onami S. Cell-to-cell heterogeneity in cortical tension specifies curvature of contact surfaces in Caenorhabditis elegans embryos. PLoS ONE. 2012;7:e30224.

Garteh J, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. New York: Springer; 2013.

Guyon I, Gunn S, Nikravesh M, Zadeh L, editors. Feature extraction: foundations and applications. New York: Springer; 2006.

Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. 2nd ed. New York: Springer; 2009.
http://cran.r-project.org/web/packages/randomForest/index.html
http://www.kegg.jp/kegg-bin/show_pathway?mmu05222
http://www.kegg.jp/kegg-bin/show_pathway?mmu05223
http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE52583
Junttila MR, de Sauvage FJ. Influence of tumour micro-environment heterogeneity on therapeutic response. Nature. 2013;501(7467):346–54.
Kiel MJ, Yilmaz ÖH, Iwashita T, Yilmaz OH, Terhorst C, Morrison SJ. Cell. 2001;121(7):1109–21.
Lopez-Otin C, Blasco MA, Partridge L, Serrano M, Kroemer G. The hallmarks of aging. Cell. 2013;153:1194–217.
Lupski JR. Genetics. Genome mosaicism – one human, multiple genomes. Science. 2013;341: 358–9.
Macaulay IC, Voet T. Single cell genomics: advances and future perspectives. PLoS Genet. 2014;10(1):e1004126.
Marte B (ed) Tumour heterogeneity. Nature. 2013;501(67):327–72.
Meacham CE, Morrison SJ. Tumour heterogeneity and cancer cell plasticity. Nature. 2013;501(7467):328–37.
Ozsolak F, Platt AR, Jones DR, Reifenberger JG, Sass LE, et al. Direct RNA sequencing. Nature. 2009;461:814–18.
Schmid J, Dussmann H, Boukes GJ, Flanagan L, Lindner AU, O'Connor CL, Rehm M, Prehn JH, Huber HJ. J Biol Chem. 2012;287(49):41546–59.
Shapiro E, Biezuner T, Linnarsson S. Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat Rev Genet. 2013;14(9):618–30.
Speicher MR. Single-cell analysis: toward the clinic. Genome Med. 2013;5:74.
Strobl C, Boulesteix A-L, Zeileis A, Hothorn T. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics. 2007;8:25.
Tang DG. Understanding cancer stem cell heterogeneity and plasticity. Cell Res. 2012;22(3):457–72.
Tomasetti C, Vogelstein B, Parmigiani G. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. Proc Natl Acad Sci U S A. 2013;110:1999–2004.
Treutlein B, Brownfield DG, Wu AR, Neff NF, et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature. 2014;509(7500):371–5.

Julián Candia is a Staff Scientist with the Center of Human Immunology at the National Institutes of Health (Bethesda, MD). He earned a degree as 'Licenciado en Física' from the University of La Plata (Argentina) in 1999, and his Ph.D. in Physics from the same university in 2004. Dr. Candia's research interests are focused on analyzing large-scale biological datasets with a variety of computational and analytical techniques at the crossroads of physics, statistics, and computer science. His particular interests are in integrating multiple types of biomedical information, including cell-biological and biophysical information, e.g. gene and microRNA expression, cell shapes, and immunophenotypes, with the goal of contributing innovative ideas to the most pressing problems in current biomedical research. For more information, see *juliancandia.com*.

Jayanth R. Banavar is Dean of the College of Computer, Mathematical, and Natural Sciences at the University of Maryland. Prior to his current appointment, Dr. Banavar served as Distinguished Professor and George A. and Margaret M. Downsbrough Department Head of Physics at Pennsylvania State University. He received a Bachelor of Science with honors and a Master of Science in physics from Bangalore University. He earned his Ph.D. in physics from the University of Pittsburgh. A fellow of the American Physical Society and the American Association for the Advancement of Science, he has more than 250 publications in refereed journals, 11 book chapters, a book he co-edited, and three patents. Much of Dr. Banavar's recent work has applied the techniques of statistical physics to solve interdisciplinary problems, explaining, for example, why biological molecules tend to curl up into helices, or to explain why coral reefs support such a rich biodiversity. Frequently, the goal has been to identify an underlying mathematical principle to provide an elegant explanation of natural phenomena.

Wolfgang Losert is Professor of Physics and Associate Dean for Research at the University of Maryland and co-founder and Director of the University of Maryland-National Cancer Institute Partnership for Cancer Technology. He received a Diploma in Applied Physics in 1995 from the Technical University of Munich, and a Doctorate in Physics from the City College of the City University of New York in 1998.

Dr. Losert's research group studies the dynamics of living systems, with a focus on new approaches to measure and understand the spatial patterns and dynamics of cells and tissues. A particular aim of the work is to better understand the physical characteristics of cancer, such as the shape, dynamics and heterogeneity of cancer cells. Through his research, Dr. Losert contributes novel approaches to harness the emerging abundance of quantitative data in biological and medical research.