Journal of
Proteomics & Bioinformatics
ISSN: 0974-276X

Lahesmaa-Korpinen et al., J Proteomics Bioinform 2011, 4:11
DOI: 10.4172/jpb.1000197

Research Article Open Access

FlowAnd: Comprehensive Computational Framework for Flow Cytometry Data Analysis

Anna-Maria Lahesmaa-Korpinen[1], Sari E. Jalkanen[2], Ping Chen[1], Erkka Valo[1], Javier Núñez-Fontarnau[1], Ville Rantanen[1], Ali Oghabian[1], Jukka Vakkila[2], Kimmo Porkka[2], Satu Mustjoki[2] and Sampsa Hautaniemi[1]*

[1]Research Programs Unit, Genome-Scale Biology and Institute of Biomedicine, Biochemistry and Developmental Biology, University of Helsinki, PO Box 63 (Haartmaninkatu 8), 00014 University of Helsinki, Finland
[2]Hematology Research Unit, Biomedicum, Division of Medicine, Helsinki University Central Hospital and University of Helsinki, Finland

Abstract

Flow cytometry is a widely used high-throughput measurement technology in basic research and diagnostics. Recently the amount of data generated from flow cytometry experiments has been increasing, both in sample numbers and the number of parameters measured per cell. These highly multivariate datasets have become too large for use with tools depending mainly on manual analysis.

We have implemented a computational framework (FlowAnd) that is designed to analyze and integrate large-scale, multi-color flow cytometry data. The tool implements methods for data importing, various transformations, several clustering algorithms for automatic clustering, visualization tools as well as straightforward statistical testing. We applied FlowAnd to a phosphoproteomics data set from 37 chronic myeloid leukemia patients treated with two kinase inhibitors. Our results indicate high concordance between automated gating using three clustering algorithms and manual gating. Analysis of more than 70 flow cytometry experiments demonstrate the utility of features in FlowAnd, such as a graphical tool for rapid validation of clustering results, in large-scale flow cytometry data analysis.

The FlowAnd framework allows accurate, fast and well documented analysis of multidimensional flow cytometry experiments. It provides several clustering algorithms for automatic gating, the possibility to add novel tools in various programming languages, such as Java, R, Python or MATLAB in an environment amenable to high-performance computing. FlowAnd can also be easily modified to comply with various marker panels and parameter settings. FlowAnd, all data and user guide are freely available under GNU General Public License at http://csbi.ltdk.helsinki.fi/flowand.

Introduction

Flow cytometry (FCM) is a high-throughput measurement technology that allows a large variety of cell level measurements from counting cell populations using cell surface markers to quantification of signaling protein levels with intracellular staining [1]. In blood cancers, in particular acute leukemia, FCM based cell counting is routinely used for disease diagnosis and measurement of minimal residual disease during follow-up [2]. An FCM capable of measuring six markers, which is a typical setting in clinical applications, can produce over 3 million data points for one patient. The need to analyze data from cohorts of patients together with the increasing numbers of FCM markers calls for computationally efficient tools to manage, analyze, visualize and integrate FCM data.

The analysis workflow from the raw FCM data to interpretable results useful for clinical decision making is complex and consists of several steps most of which are currently done manually with various software. One of the most time consuming step in the FCM data analysis is arguably gating, *i.e.*, the selection of cells of interest from the data. Software such as FlowJo (TreeStar, Ashland, OR), FCS Express (De Novo Software, Los Angeles, CA) and Cytobank [3] aim at easy-to-use manual gating. As manual gating is not practical in the analysis of FCM data from tens or hundreds or patients, several methods for automatic gating, such as SamSPECTRAL [4] and flowMeans [5], have been suggested. Gating, however, is only one step in FCM data processing and current frameworks that allow integrated analysis, such as FLAME [6], FIND [7] and flowCore [8] do not scale up to analyze millions of data points that emerge from clinical applications. Furthermore, users typically need to copy and paste results from one software to another, which makes the manual process error-prone and tedious.

We present a computational framework (FlowAnd) for comprehensive analysis of high-throughput FCM data from large cohorts of patients. FlowAnd integrates several individually published flow cytometry analysis tools to herein developed novel ones within a unified computational framework that supports parallel programming of the computationally demanding methods as well as statistical methods for downstream analyses. The FlowAnd software is thoroughly documented, actively maintained and new versions are released periodically.

Materials and methods

Data

In order to demonstrate FlowAnd we used data from a previously published experiment with intracellular phosphoprotein measurements from six-color FCM from chronic myeloid leukemia patients [9]. To test the functionality of FlowAnd and compare the implemented methods we used data for one patient and for a test of large-scale capabilities we used a large data set with 37 patient samples. All data

***Corresponding author:** Sampsa Hautaniemi, Research Programs Unit, Genome-Scale Biology and Institute of Biomedicine, Biochemistry and Developmental Biology, University of Helsinki, Finland, E-mail: sampsa.hautaniemi@helsinki.fi

Received September 22, 2011; **Accepted** November 03, 2011; **Published** November 29, 2011

Citation: Lahesmaa-Korpinen AM, Jalkanen SE, Chen P, Valo E, Núñez-Fontarnau J, et al. (2011) FlowAnd: Comprehensive Computational Framework for Flow Cytometry Data Analysis. J Proteomics Bioinform 4: 245-249. doi:10.4172/jpb.1000197

are freely available at http://csbi.ltdk.helsinki.fi/flowand. Comparison of the performance of the automatic gating methods to a golden standard of manually gated data was quantified using the F-score and correlation of the identified cell populations. The F-score is a measure of accuracy for a classification test that accounts for both precision and recall of a test:

$$F = 2 \times \frac{precision.recall}{precision + recall}$$, The range for F-score is [0,…,1].

In the multiple patient dataset there were four sample groups: healthy controls (n=7), patients at diagnosis (n=10), patients after imatinib treatment (n=10) and patients after dasatinib treatment (n=10). In order to study the differences in signaling, the intracellular signaling of the cells was induced *ex vivo* with various cytokines and a control PBS-stimulation. For baseline phosphoprotein studies fixed cells were stored without cytokine or PBS stimulation.

For each patient, due to the different stimulations, four different panels of fluorescent antibodies were used. The panels were used so that for the control stimulation all panels were measured, and for each of the three different cytokine cocktails two of the panels were measured, resulting in a total of 10 experiments and data files for each patient (Table 1 as in [9]). For the baseline study done with samples with no stimulations, the panels in Table 2 were used. Therefore when considering each panel in addition to the forward scatter and side scatter measurements, there are always a total of eight parameters measured from each cell. Each individual file typically has hundreds of thousands of cells resulting in a data set with roughly 112 million data points, which render manual analysis tedious. Details of experimental protocols can be found in [9].

Software implementation

FlowAnd runs on the freely available Anduril framework [10]. Anduril is a flexible data analysis framework that is intended for analysis and integration of data from various data sources. The Anduril framework is designed so that individual components in the system can be created with different programming languages, currently covering Java, R, Python, MATLAB and Perl. Anduril supports parallel programming and thus computationally intensive jobs can be run in parallel.

FlowAnd and user guide are freely available under GNU General Public License at http://csbi.ltdk.helsinki.fi/flowand. The main modules in FlowAnd are data import, preprocessing, cell gating, population identification and statistical analysis as shown in Figure 1. Each module consists of several algorithms. For example, the data import module accepts Flow Cytometry Standard (FCS) files. The data import functions are mainly from the flowCore package. Preprocessing module contains data transformation functions, such as logarithmic and arcsinh, as well as tools to filter outlier data points that are likely debris.

Panel ID	A488	PE	PerCP	PECy7	A647	APCH2	Conditions
1	ERK 1/2	STAT1	CD4	CD3	STAT3	CD45	Control, Stimulations A, B and C
2	STAT5	STAT1	CD4	CD3	STAT3	CD45	Control, Stimulations A
3	STAT5	STAT1	CD4	CD25	STAT3	CD45	Control, Stimulations B
4	STAT1	STAT6	CD4	CD25	STAT3	CD45	Control, Stimulations C

Table 1: Panels of fluorescence antibodies used for specific stimulation conditions. This table is adapted from [9].

Panel ID	A488	PE	PerCP	PECy7	A647	APCH7
1	ERK1/2	STAT1	CD4	CD25	STAT3	CD45
2	STAT5	STAT6	CD4	CD25	STAT3	CD45

Table 2: Panels of fluorescence antibodies used for baseline study.

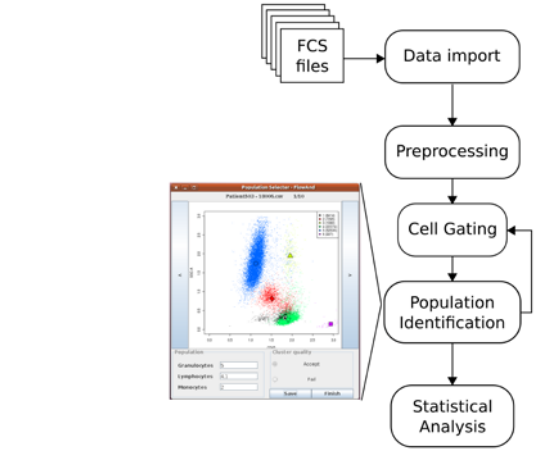

Figure 1: **FlowAnd overview:** An overview figure of the steps in a full analysis of flow cytometry data. First data are preprocessed by applying compensation and a transformation to the data. Next the cells are filtered by clustering and selecting debris clusters out from the data. The third step is to identify cell populations from the data and this is done by first clustering the data and then applying a rule based method to label the clusters into populations. Finally once the correct populations have been identified, the data can be summarized in a report.

The cell gating module currently consists of three gating algorithms: the FLAME algorithm of mixture modeling with a t-skew distribution [6], SamSPECTRAL that combines spectral clustering with a FCM tailored sampling procedure [4], and a variant of k-means clustering (flowMeans) [5]. To assist with the automation of the gating procedure, we have developed a component that allows the use of rules based on biological knowledge. The component is given *a priori* rules about the approximate locations of the cell populations as specific parameters also enabling rules relating clusters to one another. For example, it is known that the lymphocyte population has higher CD45 expression than the granulocyte and monocyte populations but a low side scatter value. Furthermore, in normal blood sample and bone marrow samples granulocytes are known to have a large number of cells and a high side scatter value with a high variance. This type of biological knowledge can be translated into rules that can be used to identify the populations.

These automatic tools do not always work perfectly and a user may want to visually inspect the clustering and population labeling manually. To allow manual intervention FlowAnd has a component that produces a graphical interface where the user can see the results of the clustering with the automatically identified cluster labels, manually correct these identifications and use them in processing the final results. The population identification can be run repeatedly for any identified population to identify subpopulations.

For final results, various statistical tests can be performed. For instance, to compare the expression of a protein in various cell population of two patient groups, or generate heatmaps for visualization of the high-dimensional FCM data.

Results

Single patient case study

We selected one patient to demonstrate a full analysis workflow and compare performances of three gating algorithms (FLAME, flowMeans and SamSPECTRAL) to manual gating results of three cell populations (granulocytes, monocytes and lymphocytes). The first step of analysis is to identify the cells from the debris, as this data contain a significant amount of debris. This is due to the fact that the samples analyzed were permeabilized cells which are required for enabling intracellular staining. Debris filtering is important because clustering methods perform much faster when only the relevant cells are used. For this filtering, data were clustered based on the scatter channels and CD-markers, a total of five variables for each file (FSC, SSC, CD45, CD4, and CD25/CD3 depending on the panel of antibodies used). The debris was filtered out automatically by selecting the clusters that had lower median values for FSC, SSC and CD45 than the average cluster median values, as is typically done when manually selecting the cell data from the raw data. This filtering step reduced the data from hundreds of thousands of cells (mean of 340,000 cells in raw data) to less than one hundred thousand cells (mean of 70,000 cells after filtering).

The smaller amount of debris-free data were used again with the clustering algorithm to obtain more precise clustering results than clustering with all of the raw data. The granulocyte, monocyte and lymphocyte populations were identified from the clusters resulting from all three methods and the results were compared to manually gated data. Figure 2 shows a representative image of clustering results by the different methods. With flowMeans and SamSPECTRAL we get relatively similar clusters and the three main populations (granulocytes, monocytes and lymphocytes) can be identified from the clustering results. The FLAME method identified too many clusters

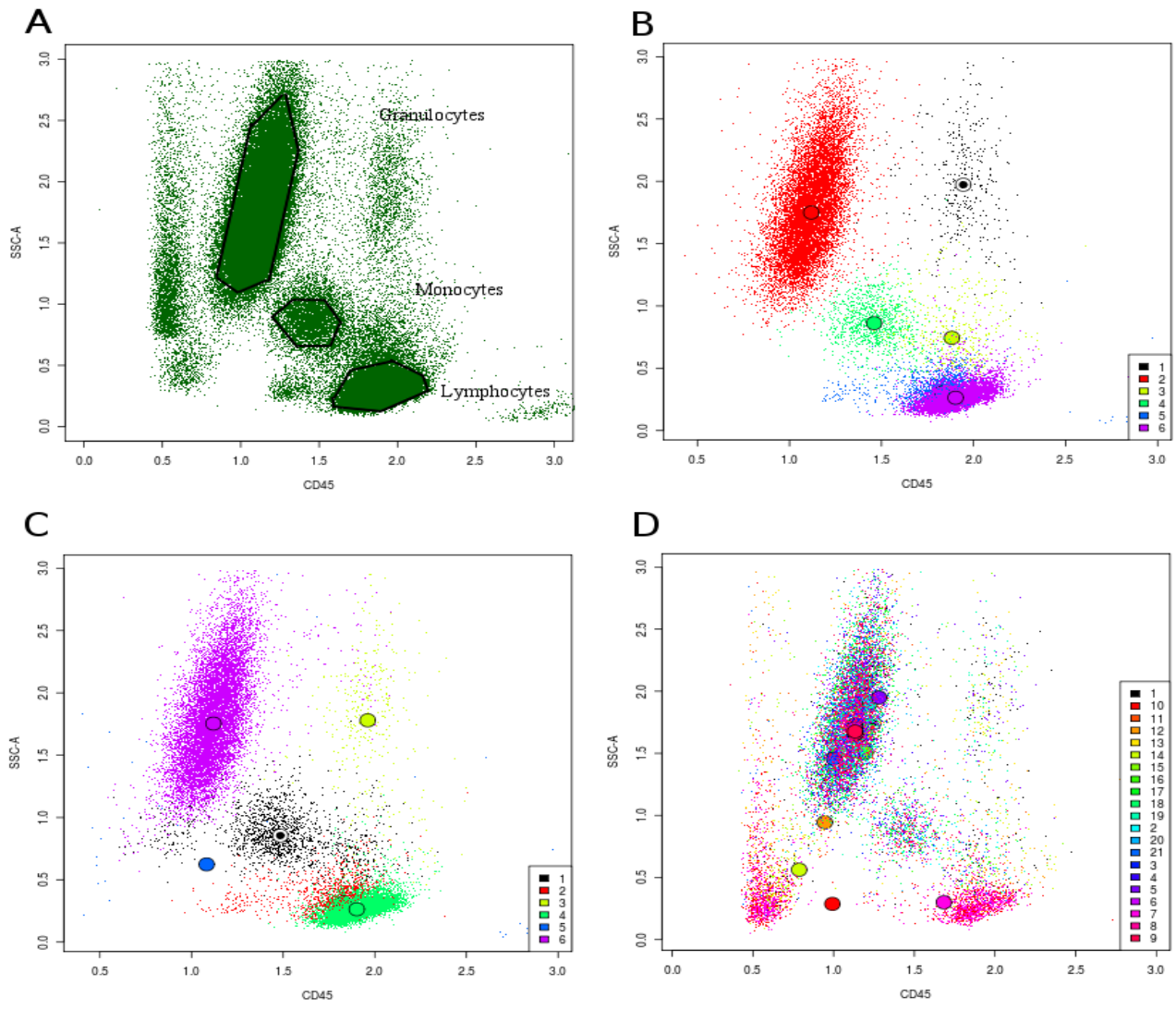

Figure 2: **Comparison of gating algorithms:** Comparison of **a)** manual gating to automatic gating with **b)** flowMeans, **c)** SamSPECTRAL and **d)** FLAME clustering. FlowMeans and SamSPECTRAL identify the three populations relatively well, while the FLAME mixture modeling with *t* skew distribution identifies too many clusters.

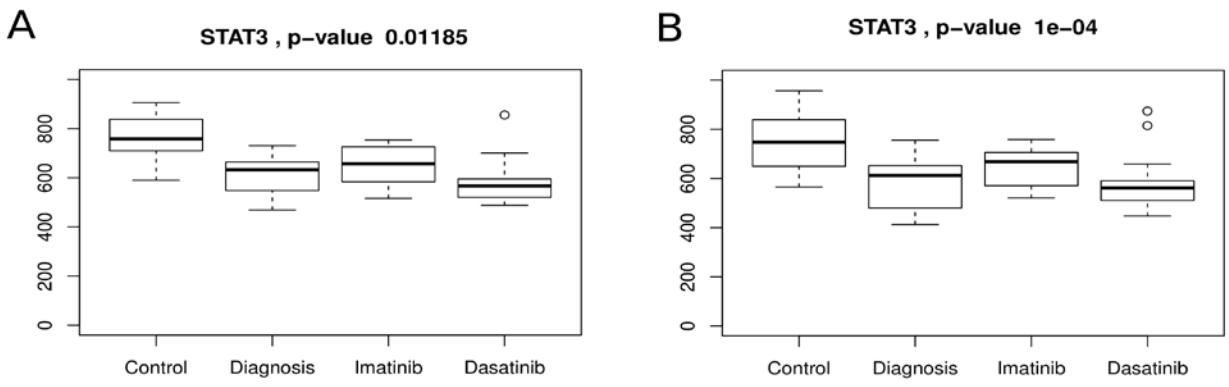

Figure 3: **Visualization of results of statistical testing: A.** The results of manual gating for STAT3 in the lymphocyte populations of 37 individuals from four patient groups, healthy controls, patients at diagnosis, patients after imatinib treatment and patients after dasatinib treatment. **B.** The replicated experiment using FlowAnd and the semi-automated analysis pipeline.

that do not separate the three main cell populations. The distribution of the cells is different with the three images because also cell and debris identification is done with different clustering results, and it can be seen that the FLAME method leaves more debris data. It is evident from the results that FCM gating algorithms can result in widely dissimilar results. Thus, the FlowAnd-type framework approach that includes several methods allows the use of the best performing method in any particular instances.

For a more detailed comparison of clustering methods we used the F-score and correlation. The most accurate method with this data was SamSPECTRAL with an F-score of 0.97 followed by flowMeans (F-score 0.91), whereas FLAME failed to give reasonable clusters (F-score 0.63, Figure 2). The correlation between manual gating and SamSPECTRAL was 0.99, whereas flowMeans and FLAME achieved 0.69 and 0.42, respectively. flowMeans was the fastest method (1.5 hours wall clock time using five parallel processes) followed by SamSPECTRAL (6h) and FLAME (40h). These results demonstrate that there are differences between the performance of gating methods, and similar results were obtained with other patient samples (data not shown). It is thus an advantage to be able to use several methods in parallel and choose the best performing one. Furthermore, FlowAnd allows the use of faster methods, such as flowMeans, for computationally demanding clustering that do not require high accuracy and of more accurate but slower methods, such as SamSPECTRAL, for identifying detailed cell populations after coarse gating.

Multiple patient case study

To demonstrate the performance of FlowAnd in the analysis of a large number of samples, we reanalyzed the data from 37 patient samples with two experiments for each patient for a total of 74 FCS experiments [9]. The aim of this study was to replicate the previous finding of a decrease in the expression of pSTAT3 in the lymphocytes of dasatinib-treated patients in comparison to healthy controls. For this, the lymphocyte populations of all patients needed to be identified and the median pSTAT3 expression of these cells calculated.

We used flowMeans for clustering due to its high speed and relatively good accuracy and also included a second round of clustering in case the automatic parametrization did not identify all of the desired cell populations. The full analysis of the data with 10 parallel processes took 8.5 hours (wall clock time), which included only 20 minutes of manual work for checking the population labels of the clustering results. Compared to manual analysis time for identification of three populations from 74 FCM-experiments, this is a significant decrease in time. The FLAME webservice or standalone versions were not able to handle these data.

We compared the expression of phosphoproteins obtained using FlowAnd in comparison to manually gated lymphocytes in [9]. We used median values of pSTAT3 expression in lymphocytes from manually gated and computationally gated data. Both data sets were analyzed with a Kruskall-Wallis test with a null hypothesis of equal population medians for the four patient groups: healthy controls (n=7), patients at diagnosis (n=10), patients after imatinib treatment (n=10) and patients after dasatinib treatment (n=10). A p-value of less than or equal to 0.05 was considered significant. The manually gated lymphocyte data are plotted in Figure 3A and the FlowAnd gated data in Figure 3B and both methods gave similar results showing that there is a difference in the expression of pSTAT3 and that the difference was between the control and dasatinib treated individuals. Similar analyses were done with FlowAnd for other populations and markers (data not shown).

Discussion

FlowAnd is designed to allow the analysis of large-scale FCM experiments with tens to hundreds of patients and multiple FCM experiments for each patient. Our objective was to create a framework that is scalable, enable the use of different clustering algorithms, and provide an environment where analysis is straightforward, repeatable, and rapid in comparison to manual analysis. These features were demonstrated with two case studies comprising of one and 37 leukemia patients. When using Cytobank or FlowJo, the manual gating is a time consuming process and for all downstream analysis, the values must be copied from the original software to another statistical software. FlowAnd is implemented in Anduril, which allows taking advantage of tools for multivariate statistics, such as Weka, MATLAB and R, in a

unified framework.

As complexity and numbers of FCM analyses are increasing in research and diagnostic laboratories, there is a need for computational frameworks, such as FlowAnd, that allow accurate, fast and well documented analysis of multidimensional FCM experiments. The results of these case studies demonstrate that FlowAnd is able to efficiently process large-scale FCM data as well as integrate analysis tools into a coherent framework. FlowAnd can be easily modified to comply with various marker panels and parameter settings.

Acknowledgements

Funding: Academy of Finland (projects 125826, 136181), Sigrid Jusélius Foundation, Finnish Cancer Associations, Helsinki Biomedical Graduate School and Biocentrum Helsinki, K.A. Johanssen Foundation and Finnish Association of Haematology.

References

1. Chattopadhyay PK, Hogerkorp CM, Roederer M (2008) A chromatic explosion: the development and future of multiparameter flow cytometry. *Immunology* 125:441–449.
2. Peters JM, Ansari MQ (2011) Multiparameter flow cytometry in the diagnosis and management of acute leukemia. *Arch Pathol Lab Med* 135:44–54.
3. Kotecha N, Krutzik PO, Irish JM (2010) Web-based analysis and publication of flow cytometry experiments. *Curr Protoc Cytom* 10:Unit10.17.
4. Zare H, Shooshtari P, Gupta A, Brinkman RR (2010) Data reduction for spectral clustering to analyze high throughput flow cytometry data. *BMC Bioinformatics* 2010, 11:403.
5. Aghaeepour N, Nikolic R, Hoos HH, Brinkman RR (2011) Rapid cell population identification in flow cytometry data. *Cytometry A* 79:6–13.
6. Pyne S, Hu X, Wang K, Rossin E, Lin TI, et al. (2009) Automated high-dimensional flow cytometric data analysis. *Proc Natl Acad Sci U S A* 106:8519–8524.
7. Dabdoub SM, Ray WC, Justice SS (2011) FIND: a new software tool and development platform for enhanced multicolor flow analysis. *BMC Bioinformatics* 12:145.
8. Hahne F, LeMeur N, Brinkman RR, Ellis B, Haaland P, et al. (2009) flowCore: a Bioconductor package for high throughput flow cytometry. *BMC Bioinformatics* 10:106.
9. Jalkanen SE, Vakkila J, Kreutzman A, Nieminen JK, Porkka K, et al. (2011) Poor cytokine-induced phosphorylation in chronic myeloid leukemia patients at diagnosis is effectively reversed by tyrosine kinase inhibitor therapy. *Exp Hematol* 39:102–113.
10. Ovaska K, Laakso M, Haapa-Paananen S, Louhimo R, Chen P, et al. (2010) Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. *Genome Med* 2:65