

Resolving and classifying haematopoietic bone-marrow cell populations by multi-dimensional analysis of flow-cytometry data

Eli Zamir,² Benjamin Geiger,² Nir Cohen,³ Zvi Kam² and Ben-Zion Katz¹

¹The Haematology Institute, Tel-Aviv Sourasky Medical Centre, Tel-Aviv, Israel, ²The Department of Molecular Cell Biology, The Weizmann Institute of Science, Rehovot, Israel, and ³The Department of Orthopaedics, Tel-Aviv Sourasky Medical Centre, Tel-Aviv, Israel

Summary

The study of normal or malignant haematopoiesis requires the analysis of heterogeneous cell populations using multiple morphological and molecular criteria. Flow cytometry has the capacity to acquire multi-parameter information of large haematopoietic cell populations, utilizing various combinations of >200 molecular markers (clusters of differentiation, CD). However, current flow cytometry analyses are based on serial gating of two-parametric scatter plots – a process that is inherently incapable to discriminate all subgroups of cells in the data. Here we studied the cellular diversity of normal bone marrows (BM) using multi-dimensional cluster analysis of six-parametric flow cytometry data (four CD, forward scatter and side scatter), focusing mainly on the myeloid lineage. Twenty-three subclasses of cells were resolved, many of them inseparable even when examined in all possible two-parametric scatter plots. The multi-dimensional analysis could distinguish the haematopoietic progenitors according to International Society of Haematotherapy and Graft Engineering criteria from other types of immature cells. Based on the defined clusters, we designed a classifier that assigns BM cells in samples to subclasses based on robust six-dimensional position and extended shape. The analysis presented here can manage successfully both the increasing numbers of haematopoietic cellular markers and sample heterogeneity. This should enhance the ability to study normal haematopoiesis, and to identify and monitor haematopoietic disorders.

Keywords: bone marrow, flow cytometry, leukaemia, leucocytes, myeloid.

Received 10 November 2004; accepted for publication 3 February 2005

Correspondence: Katz Ben-Zion, PhD, The Haematology Institute, Tel-Aviv Sourasky Medical Centre, 6 Weizman St., Tel-Aviv 64239, Israel. E-mail: bkatz@tasmc.health.gov.il

Haematopoiesis is a complex developmental process whereby multiple cell lineages residing in the bone marrow (BM), peripheral blood and lymphatic organs, undergo differentiation and proliferation. These cells originate from a small number of stem cells with self-renewing potential, which give rise to a mixture of cells with different levels of maturation. Many of these cell types can be identified using lineage-specific and differentiation-specific antibodies (Borowitz *et al*, 1997). In view of its central physiological importance, haematopoiesis is a tightly regulated process, and any deviation from normal development patterns may result in severe disorders. Particularly devastating are different neoplasms such as leukaemias, lymphomas and myeloproliferative disorders, where cells belonging to a particular lineage and stage of differentiation become malignant. Thus, BM monitoring is essential for the

detection, diagnosis and follow-up of haematopoietic pathologies and their therapeutic treatments.

Flow cytometry is a powerful and widespread approach for studying normal and aberrant haematopoiesis (Weir & Borowitz, 2001). So far, >200 haematopoietic membrane and cytoplasmic markers have been classified as 'clusters of differentiation' (CD), and their expression patterns are used to characterize the differentiation stages of cells belonging to various haematopoietic lineages (e.g. lymphoid, myeloid and erythroid). Advanced Fluorescence Activated Cell Sorter (FACS) instruments can simultaneously detect four or more CD markers, each labelled with a different fluorophore (De Rosa *et al*, 2001, 2003, De Rosa & Roederer, 2001). In addition, two non-fluorescent physical parameters are routinely recorded by flow cytometry, namely forward scatter (FSC),

which is a measure of cell size, and side scatter (SSC), which evaluates cell granularity (Weir & Borowitz, 2001). The FACS acquisition capabilities are continuously increasing, currently extending beyond 10 colours, and expanding the gap between data generation and its analysis (De Rosa *et al*, 2001, 2003, De Rosa & Roederer, 2001).

Since the introduction of flow cytometry for research and clinical haematology in the early 70s, numerous studies have addressed the issue of 'immuno-phenotyping' of BM cell populations (Brown *et al*, 1975; Cantor *et al*, 1975). The underlying rationale is that detailed lineage- and differentiation-specific cell typing is crucial for accurate clinical diagnosis and selection of an optimal therapeutic approach. However, the commonly applied analysis of the multi-parametric information is rather incomprehensive and complex (Jennings & Foon, 1997; Hrusak & Porwit-MacDonald, 2002). Hence, morphology, cytogenetics and cytochemistry are considered mandatory for the diagnosis and prognosis of many disorders such as acute myeloid leukaemia (AML), while only a minor role is attributed to flow cytometry (Appelbaum *et al*, 2000; Vardiman *et al*, 2002).

Part of the problem in defining a high precision, flow cytometric-based lineage mapping is attributable to the heterogeneity and variability of both normal and pathological haematopoietic cell populations, rendering difficult the definition of a 'normal' BM template in terms of the measured parameters. We have developed algorithms that address these inherent cellular properties and applied them to multi-parameter flow cytometry of clinical data.

A second problem is that the multi-parametric flow cytometric data must be treated by a truly multi-dimensional analysis in order to fully resolve the different cell types in the data. Multi-parametric analysis procedures, including computerized algorithmic procedures and interactive visualization tools, have long been under development for data sorting in many fields, in natural and social sciences as well as for economic purposes (Blatt *et al*, 1996; Richard *et al*, 1999). The steady increase in computer power enables the application of such tools for large databases, including multi-parameter flow cytometric measurements (Beckman *et al*, 1995; Boddy *et al*, 2001). In bioinformatics research, for example, the huge volume of available genomic or proteomic data synergized with the development of powerful algorithms for the search of gene sequence homologies, including in prognostic evaluation of leukaemias (Alon *et al*, 1999; Bullinger *et al*, 2004; Valk *et al*, 2004). A powerful approach to characterize pattern similarity in multi-dimensional spaces is cluster analysis. The various mathematical algorithms for clustering multi-parameter data are based on grouping data points that are densely packed in the parameter space (a space in which each dimension corresponds to one of the measured parameters). Thus, for example, a typical flow cytometric scatter plot, presenting cells based on only two of the measured flow cytometric parameters, is usually 'clustered' visually, enabling the definition of specific cell populations. However, the

obvious limitation of this visual approach is that only a fraction of the data is being observed at any time, and, for example, cells that seem similar in such two-dimensional presentations (which is a projection of the multidimensional distribution), may actually have very different values for the other measured parameters. Moreover, even in two-dimensional projections, the different cell lineages cluster in widely varying shapes, ranging from compact to highly elongated ones. The clustering approach described here addresses these issues by simultaneous examination of all six parameters, and redefining the multi-parametric distance to clusters based on their shape. Previous studies have utilized cluster analysis methodologies to screen large numbers of flow cytometry-based data files for patterns of disease (but not at the individual BM level), or to analyse small numbers of parameters in specific diseases (e.g. chronic lymphocytic leukaemia) (Valet & Hoffkes, 1997; Wells *et al*, 2003).

In the present study, we developed a new approach for the classification of subpopulations of BM-derived haematopoietic cells, based on a panel of four myeloid markers and light scattering data. Using this approach, we have resolved 23 cell subclasses in normal BM, 13 of which belong to the myeloid/monocytic lineage. This cell classification scheme was reproducible among normal individuals, thereby forming a normal BM population template. In addition, focusing on CD34 expressing cells, we identified a cluster representing the 'classical' CD34⁺ haematopoietic progenitors according to International Society of Haematotherapy and Graft Engineering (ISHAGE) criteria, as well as six more subclasses that expressed this marker of immature cells. Some of these cell types represent immature myeloid cells based on their marker pattern, and classification of myeloid cell lines and primary AML samples. Their abundance and properties in normal BM are described. The ability to analyse simultaneously the six (or in fact any number of) parameters acquired by advanced flow cytometry enhances significantly the analysis of physiological and pathological haematopoiesis, and provides a powerful basis for the identification, characterization, and diagnosis of haematological disorders.

Materials and methods

Preparation of BM samples

Specimen collection was approved by the Institutional Review Board at Tel-Aviv Sourasky Medical Centre. Bone fragments from orthopaedic surgery were washed to discard peripheral blood contamination, and BM was extracted by *in vitro* aspiration. The BM samples from normal bone fragments or primary AML BM aspirates were counted utilizing a Coulter MaxM counter (Coulter Corp., Fullerton, CA, USA). As confirmed by blood counts (platelets and red blood cells), normal BM samples had <10% contamination with peripheral blood (not shown), similar to the first millilitre of BM aspirated from healthy donors (Batinic *et al*, 1990). The BM

Table I. CD markers used in the present study.

CD	Protein characteristic	Expression specificity
CD13	A 150 kDa membrane associated aminopeptidase	Myeloid and monocytes differentiation [Riemann <i>et al</i> (1999); Dybkaer <i>et al</i> (2001)]
CD33	A 67 kDa glycosylated transmembrane protein	Pan-myeloid/monocytic marker [Mingari <i>et al</i> (2001)]
CD34	A 116 kDa heavily glycosylated transmembrane protein	Immature haematopoietic stem/progenitor cells [Engelhardt <i>et al</i> (2002)]
CD45	A 220 kDa transmembrane protein tyrosine phosphatase	All nucleated haematopoietic cells [Penninger <i>et al</i> (2001)]

The characteristics and expression specificities of the CD markers used in this study. These markers are expressed by immature and well-differentiated monocytic/myeloid cells, and are commonly utilized to analyse myeloid malignancies (Braylan *et al*, 2001).

samples were labelled with a panel of four membrane markers (Table I), which are commonly utilized in flow cytometric analysis of AML (Braylan *et al*, 2001). The specificity of the markers according to the normal haematopoietic lineages is detailed in Table I. From each sample, 3×10^4 events were acquired at a rate of 150–300 events per second by the FACSCalibur instrument and CellQuest software (Becton-Dickinson, San Jose, CA, USA).

Cell lines and peripheral blood neutrophils isolation

The U937 and K562 human myeloid cell lines were obtained from the American Type Culture Collection (Manassas, VA, USA). THP-1 human myelomonocytic leukaemia cell line was kindly provided by Isaac P. Witz (Tel-Aviv University, Tel-Aviv, Israel). The cell lines were cultured in Roswell Park Memorial Institute (RPMI) medium supplemented with 1 mmol/l glutamine, 50 µg/ml streptomycin, 50 U/ml penicillin, and 10% heat-inactivated bovine serum (Biological Industries, Beit-Ha'Emek, Israel) at 37°C, in a 5% CO₂ humidified incubator. Primary human neutrophils were obtained from healthy donors. Briefly, heparinized peripheral blood from healthy donors was mixed 1:1 with sterile phosphate-buffered saline (PBS), and layered on Ficoll-Paque (Amersham, Arlington Heights, IL, USA), followed by centrifugation at 320 g for 20 min. Neutrophils were collected, washed and suspended in RPMI medium. Nearly 95% neutrophils purification was confirmed by morphological examination and conventional flow cytometric analysis.

Data analysis

The presented analysis of flow cytometric data included two steps. In the first step, a set of normal, reference, BM samples was subjected to multi-dimensional clustering analysis to resolve subclasses of cells. This step needed to be carried only once, for constructing the classifier, and it incorporated both objective information about six-parametric similarities and possibly additional expert knowledge about the specific system. In the second step, new, tested, BM samples were subjected to

multi-dimensional cell classifications, to assign each cell in the sample to one of the previously defined subclasses. This step is fast, automatic, and can be routinely applied for diagnostics of BM samples. The substeps of the analysis are described below, based on the following flowchart:

Step one. Definition of subclasses of cells in normal BM samples:

- 1 Pooling flow cytometric data of several normal samples (see 'Data organization and normalization').
- 2 Over-splitting the data to small and compact clusters of cells based on six-parametric similarities (see 'Top-down clustering').
- 3 Merging the compact clusters, based on bottom-up hierarchical clustering and expert judgment to ellipsoidal shaped clusters, termed here as subclasses of BM cells (see 'Bottom-up merging').

Step two. Classification of cells in test BM samples to subclasses.

Data organization and normalization

Six normal BM samples were labelled, and acquired by flow cytometry (3×10^4 cells in each sample). The flow cytometric data files were converted to tab-delimited matrix format using the FLOWJO software (Tree Star Inc., OR, USA), and imported to the MATLAB environment (The MathWorks Inc., MA, USA), in which all the data analysis was performed. The complete MATLAB function codes are available as supplementary material that can be downloaded from the web at http://www.weizmann.ac.il/mcb/ZviKam_BJHos (see *Supplementary material*: Appendix S1). The six converted files were merged to one matrix, $P_i(j)$, which consisted of 1.8×10^5 rows, corresponding to cells (denoted by index j), and six columns (denoted by index i) corresponding to the measured flow cytometric parameters for each cell (see *Supplementary material*: Appendix S1: MATLAB function 'merge_text_files.m'). As each of the flow cytometric parameters has different units and range of numerical values, it is important to equalize their contributions to the six-dimensional Euclidean

distances in the parameter space. Therefore, each parameter i [i.e. each column in the matrix $P_i(j)$] was normalized to have a mean equal to zero and a standard deviation of one:

$$P_i^{\text{Norm}}(j) = \frac{\{P_i(j) - \text{average}(P_i)\}}{\text{standard-deviation}(P_i)} \quad (1)$$

Top-down clustering

The rows of the normalized matrix were then subjected to top-down binary clustering based on deterministic annealing algorithm (Alon *et al*, 1999) (Appendix S1: MATLAB function 'find_clusters.m'). Briefly, this algorithm split the data set into two clusters that formed two subsets with minimal multidimensional distances to their centres of mass. The binary splitting was applied recursively to split each cluster into two, until the requested number of splits was performed. The advantage of this algorithm is its speed and ability to handle large data sets. The disadvantage is that the clusters may not evenly present an optimal global clustering of the data points. Notably, at any level of split, some cluster branches may be over-divided, while others may contain clusters that require further splitting. We therefore applied seven sequential splits, yielding 128 clusters, which clearly presented over-splitting of the data.

Bottom-up merging

The 128 clusters of the previous step were then merged using a modified hierarchical clustering algorithm (Appendix S1: MATLAB function 'merge_clusters.m'), in which distances between all of the clusters were calculated, the two closest clusters were then merged into one, and so on recursively, until all of the clusters were merged. In order to take into consideration the typically elongated shapes of clusters, we defined the distance between them based on their best-fitted six-dimensional ellipsoids. For fitting an ellipsoid for a cluster k , its centre of mass, C_i^k , and its second moment matrix, M_{i_1, i_2}^k , were calculated from all cells, j^k , belonging to the cluster:

$$C_i^k = \text{average}\{[P_i^{\text{Norm}}(j^k)]\} \quad (2a)$$

$$M_{i_1, i_2}^k = \text{average}\{[P_{i_1}^{\text{Norm}}(j^k) - C_{i_1}^k] \times [P_{i_2}^{\text{Norm}}(j^k) - C_{i_2}^k]\} \quad (2b)$$

The distance $D^k(j)$ between a cell j (presented by six parameters $P_i^{\text{Norm}}(j)$ for $i = 1, 2, \dots, 6$) and the cluster k was then defined as:

$$D^k(j) = \Sigma\{[P_{i_1}^{\text{Norm}}(j) - C_{i_1}^k] \times [M_{i_1, i_2}^k]^{-1} \times [P_{i_2}^{\text{Norm}}(j) - C_{i_2}^k]\} \quad (3)$$

where the summation is over the six parameters (i_1 and $i_2 = 1, 2, \dots, 6$), and M^{-1} is the inverse second-moment matrix. This distance (known as Mahalanobi distance) can be intuitively understood as the expansion factor multiplying the best fitted ellipsoid (while keeping its orientation and axes ratios) so that it will pass through the point representing the cell in the parameters space. Thus, Mahalanobi distances from an ellipsoid will be shorter for points located along its long axis

than for points, which have the same Euclidean distance from the ellipsoid centre but are along its short axis. This enabled resolving close by elongated clusters, typical of the multi-dimensional distribution pattern of flow-cytometric-measured cell populations. For calculating the distance between two clusters, q and w , the sum of the Mahalanobi distances of all cells of cluster q from the ellipsoid fitted to cluster w , D_{tot}^{qw} , and of all cells of cluster w from the ellipsoid of cluster q , D_{tot}^{wq} , were calculated. The distance between clusters, q and w , as used by the hierarchical clustering, was then defined as $D_{\text{tot}}^{qw} + D_{\text{tot}}^{wq}$. The sequential merging process is presented by a dendrogram showing the order by which the 128 clusters had been merged (Fig 1C), where each node presenting the merged cluster is connected to the two original clusters by two edges, whose length denotes the distance between the clusters.

Having the huge data set of 1.8×10^5 cells objectively represented by 128 over-divided clusters (Fig 1B), and the hierarchy of merging them to one dendrogram (Fig 1C), it became possible to identify meaningful merged-clusters (nodes in the dendrogram, Fig 1C) presenting distinguishable subclasses. For that, each merging event (i.e. each of the nodes along the dendrogram) was examined at all possible two-dimensional projections (scatter plots) of the six-dimensional flow cytometric data. In these plots, the two involved clusters, and the rest of the cells, were marked with three different colours. If the clusters overlapped in various projections, the merging was approved. Otherwise, the two clusters were kept as two separate subclasses. This decision was semi-subjective, involved expert judgment and possibly additional knowledge about the specific system. Yet, it was objectively based on the clusters resolved by the top-down clustering and the hierarchy of their down-top merging. Hence, the multi-dimensional quality of the whole analysis was maintained. Thus, 23 subclasses were defined, and were further mapped to the major BM cell classes (immature cells, monocytes, granulocytes, nucleated red blood cells (NRBC) and lymphocytes) based on known flow cytometric parameters (Fig 1C).

Classifying BM cells in test samples

Six-dimensional ellipsoids were fitted to each of the 23 defined subclasses in the reference data (1.8×10^5 cells), as described above (Equations 2a and 2b) (Appendix S1: MATLAB function 'store_parms.m'). Test BM samples were normalized as in Equation 1, but using the average and the standard deviation of the reference data rather than of the test data. This was to make the alignment between the test data and the reference data, in the parameter space, insensitive to potential large fluctuations in the cell-type distribution of the test data (e.g. a significant decrease in the abundance of one cell type). Mahalanobi distances (Equation 3) from each cell in the tested data to each of the 23 six-dimensional ellipsoids (presenting cell subclasses) were calculated. Each cell was assigned to the subclass with which it had the shortest distance. (Appendix S1: MATLAB function 'analyze_test_sample.m').

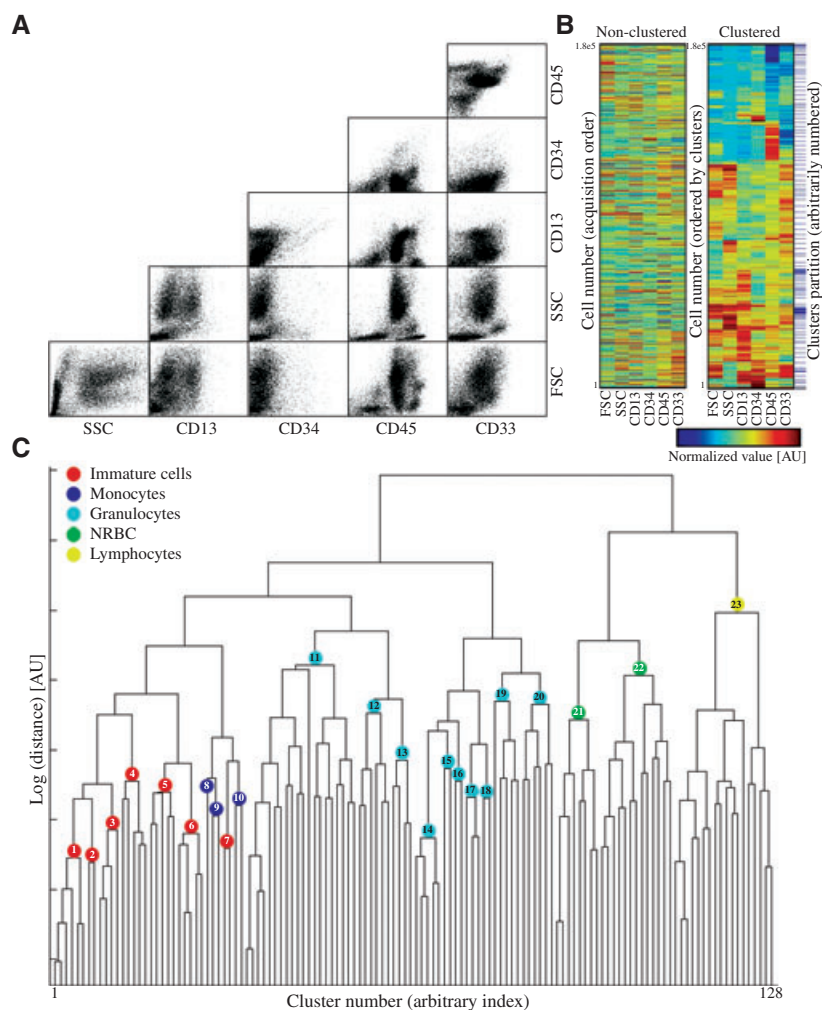


Fig 1. The analytical steps in the identification of BM cell populations. The BM samples were labelled for CD13, CD33, CD45 and CD34. For each sample, 3×10^4 cells were acquired by FACS, and six samples were grouped to a matrix with 1.8×10^5 rows (corresponding to cells) and six columns (corresponding to the six flow cytometric-measured parameters: FSC, SSC, CD13, CD33, CD45 and CD34). (A) The data is presented in all 15 possible two-parameter projections by scatter plots. For clarity, the scatter plots show a random sample of 10% from the total number of cells. (B) ‘Non-clustered’: a colour-scale presentation of the normalized data matrix in which the rows are ordered as acquired by the FACS. ‘Clustered’: the data matrix is reordered by top-down clustering, such that the rows are grouped according to their cluster. The horizontal lines indicate the partitions of the matrix to the 128 clusters. (C) A dendrogram presenting the hierarchical merging of the 128 clusters. Numbered circles indicate the nodes that stopped the merging and defined the 23 subclasses of BM cells. The colours of the circles assign the subclasses to the five main classes of BM lineages.

Each cell was also assigned to a major class of BM cells (immature cells, monocytes, granulocytes, NRBC or lymphocytes), based on the mapping of subclasses to classes (Fig 1C).

Results

Six normal BM samples were acquired by flow cytometry, measuring six parameters (FSC, SSC, CD13, CD33, CD34 and CD45) for each cell. Figure 1A shows cells of the pool of these samples at all the 15 possible two-parametric projections. Each of these two-dimensional projections alone separated the cells into a small number of distinguishable clusters (up to four). This was because two populations could be similar in some parameters but different in others. In order to resolve all distinguishable cell populations, the flow cytometric data was

analysed in the six-dimensional space, in which each dimension corresponds to one of the measured flow cytometric parameters. We applied the efficient top-down binary clustering algorithm performing six sequential splits, to yield 128 (2^7) clusters, each containing cells with similar values for all six parameters (Fig 1B, compare ‘non-clustered’ and ‘clustered’). Apparently, many of these clusters reflected over-splitting of a contiguous parent cluster. However, over-splitting was important for resolving small clusters which otherwise would have been absorbed inside big and elongated clusters. This clustering step also reduced the complexity of the data from 1.8×10^5 cells to 128 well-defined compact clusters, which could be readily examined in great detail by the hierarchical merging process (Fig 1C, and ‘Bottom-up merging’ in *Materials and methods* section). This process resulted in the identification of

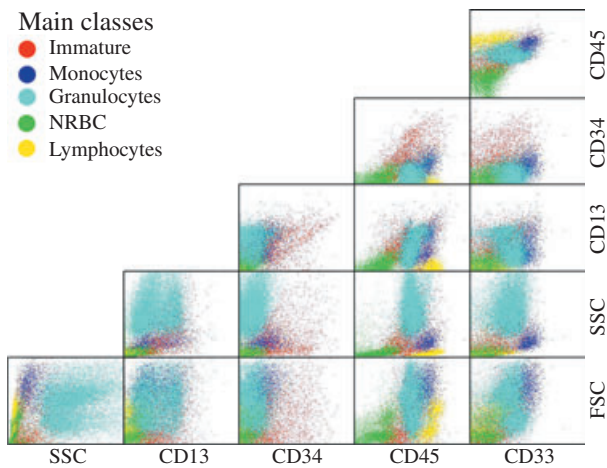


Fig 2. Characterization of the five main classes of BM cells according to the measured FSC, SSC, CD13, CD33, CD45 and CD34 values. The scatter plots show the pool of six BM samples, used for defining the cell subclasses, in all of the 15 possible two-parameter projections. Each dot in the scatter plots represents one cell and is coloured according to the main class the cell is assigned to. For clarity the plot contains a random sampling of 10% from the total number of acquired cells.

23 final clusters, which were consequently defined as cell subclasses (Fig 1C). Based on their distribution within the commonly recognized flow cytometric cell lineages and their flow cytometric values (mainly SSC and CD45), these populations could be mapped to major five cell classes composing a normal BM, namely: immature cells, monocytes, granulocytes, NRBC and lymphocytes (Figs 1C and 2). Other than one exception (subclass number 7), the mapping of the 23 subclasses to the major five classes was consistent with the dendrogram topology (Fig 1C). This indicates high correlation between the calculated Mahalanobi distances and the independently established major lineages.

Following the identification of classes and subclasses in the six-parametric space, their flow cytometric characteristics could be visualized as distributions in all possible two-parametric combinations (Figs 2 and 3), or as signatures defined by their average values (Fig 4). Non-myeloid NRBC cells did not express any of the surface markers except for a low level of CD45 (Figs 2, 3F and 4). Lymphocytes were characterized by high levels of CD45, no expression of myeloid or CD34 markers, and low SSC (Fig 2 and 4). The distribution of the major five classes in the SSC/CD45 two-parametric projection complies with the typically known patterns (Fig 2). However, as seen in this projection, these populations were not well separated, and did not form visually compact clusters with sharp borders. Most importantly, cells in the border zones between the populations in the SSC/CD45 projection were not arbitrarily assigned to one of the classes, but rather assigned based on their other flow cytometric parameters. This indicates that the SSC/CD45 projection, or any other two-parametric projection, by itself is not sufficient for a clear separation even between the five major classes, and

that multi-dimensional analysis is really essential for good classification of the cells.

In order to assign BM cells to subclasses in test samples, we designed a classifier for the 23 subclasses, considering their signatures, shapes and distributions in the six-dimensional space. As can be appreciated from the two-parametric projections (Fig 3), the shape of many cell populations is elongated rather than compact. Therefore, it would be incorrect to use Euclidian distance between a tested cell and the cluster centre-of-mass as a criterion for classification. Instead, our classifier uses the six-dimensional ellipsoid fitted to each of the 23 subclasses, and calculates the Mahalanobi distances from each six-dimensional point (representing a cell in the tested sample) to each ellipsoid. Based on that distance calculation, each cell is finally assigned to its closest ellipsoid ('Classifying BM cells in test samples' section in the *Materials and methods*).

Using the classifier, we examined the patterns of distribution of 10 individual normal BM samples. The NRBC and lymphocyte fractions obtained from the nucleated cell population in normal BMs were $19.3 \pm 6.5\%$ and $8.7 \pm 3.0\%$, respectively (Fig 5). These numbers agreed well with the values of c. 1:4 NRBC to all nucleated cell (range 13–40.1%), and 13.1% from all nucleated cells for lymphocytes (range 6–20%), reported for morphological evaluation of normal adult BM aspirates (Bain, 1996). As our antibody panel did not include specific lymphoid and/or erythroid markers, many clusters that belong to these lineages (Fig 1C) were designated as non-informative at this stage. However, the lineage identities were confirmed as lymphoid and erythroid, using lymphoid (CD3, CD19) and erythroid (Glycophorin A, CD71) flow cytometry markers (data not given).

Granulocytes are highly granulated, express myeloid markers, and CD45^{dim} (Figs 2 and 4). However, multi-parameter analysis classified 10 different granulocyte subclasses expressing variable levels of CD13 and CD33 (Figs 1C, 3A,B and 4). Total granulocytes comprised 50–60% of the nucleated BM cells (Fig 5). Interestingly, nearly 50% of all the BM granulocytes comprised a single subclass, number 11, (Fig 6C) characterized by high CD13 and low CD33 levels (Figs 3A and 4). As the predominant differentiation products of the myeloid lineage are peripheral blood neutrophils, we isolated normal peripheral blood neutrophils and examined their distribution within the BM subclasses. As given in Fig 6C, mature neutrophils were classified to subclass 11, indicating that this subclass is compatible with end-point neutrophil differentiation. These cells may also include genuine peripheral blood neutrophils that often contaminate BM aspirates. No other significant granulocyte subclasses were identified in peripheral blood granulocytes, indicating that these subclasses are apparently non-terminally differentiated granulocytes, yet to be defined. Interestingly, some of the neutrophils were classified as immature subclass number 4. As given in Fig 4, these cells were characterized by very low levels of CD34, low levels of CD13 and no CD33 expression. This may indicate that

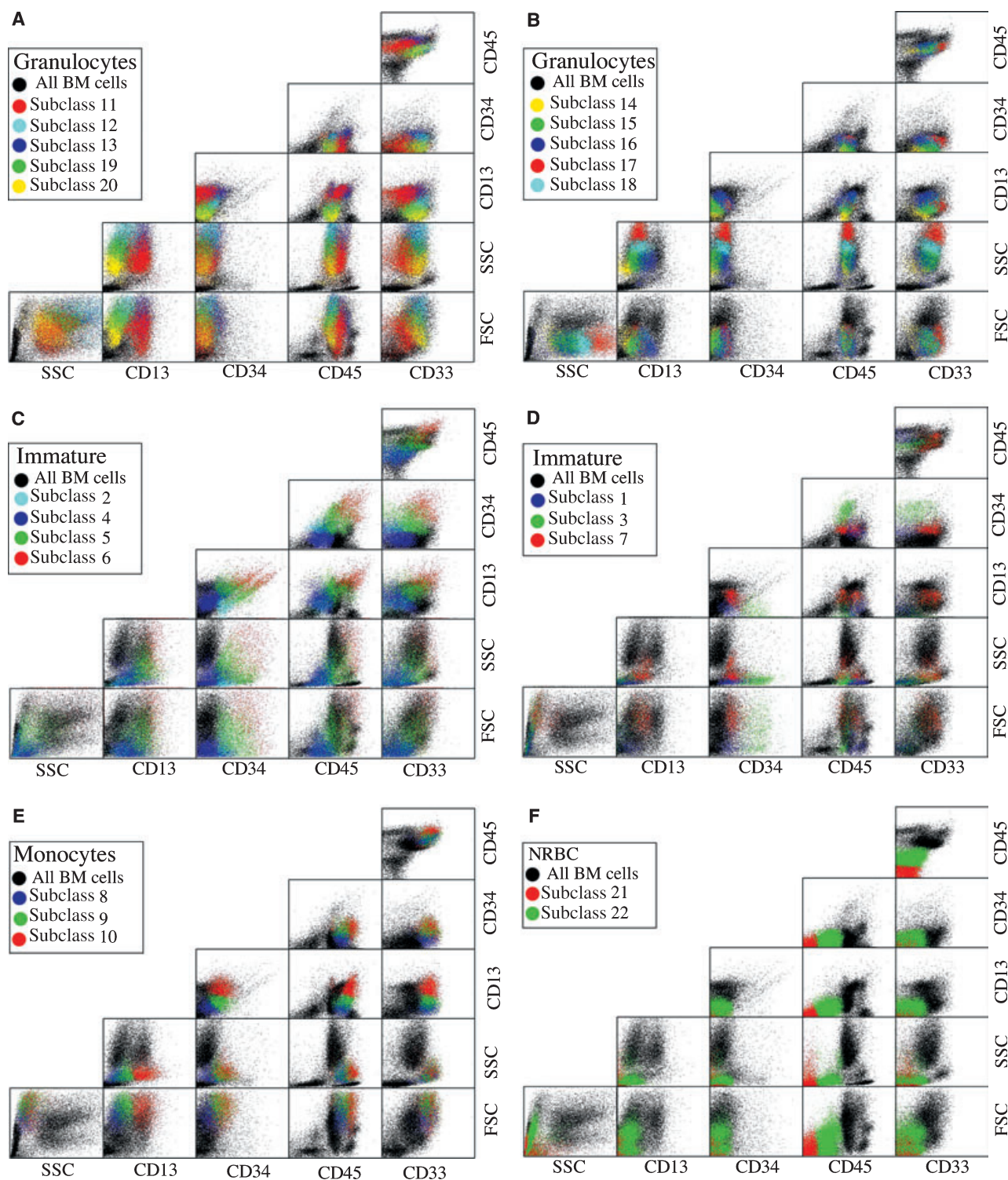


Fig 3. The distribution patterns for the 23 subclasses of BM cells according to the FSC, SSC, CD13, CD33, CD45 and CD34 values. The scatter plots show all 15 possible two-parameter projections for the pool of six BM samples used for defining the subclasses. Each dot in the scatter plots represents one cell. In each of the panels, A–F, one of the main lineage cells is coloured according to the assigned subclass (note that because of the limited number of distinguishable colours, the colour assignments do not correspond to those in Figs 1 and 2). All cells that do not belong to the lineage are marked black. (A) The more abundant granulocytes, with 5 subclasses. (B) The less abundant granulocytes, with five subclasses. (C) and (D) Immature cells, where the seven subclasses are divided to two panels for the clarity of presentation. (E) Monocytes, with three subclasses. (F) NRBC, with two subclasses. Lymphocytes, that were represented by only one subclass, are not given here (see Fig 2).

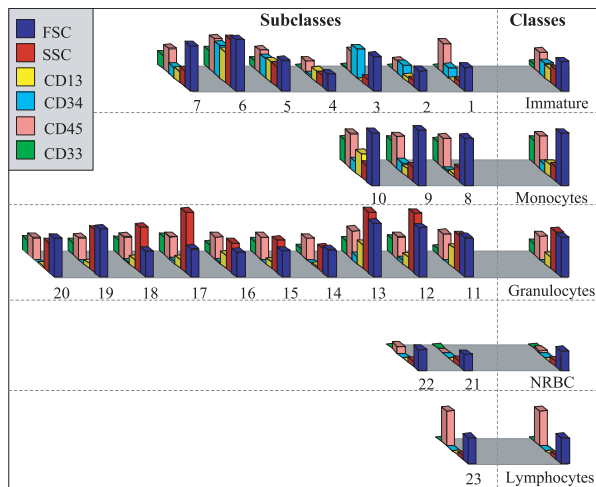


Fig 4. The six-dimensional signatures of the 23 BM subclasses found for the 1.8×10^5 -cells data set. The average, non-normalized, values of the six measured flow cytometric parameters, for each cell class and subclass, were calculated. The average values of CD13, CD33, CD45 and CD34 in the isotype controls were subtracted, and some that resulted in slightly negative values were set to zero. The six-parametric signatures are presented using bar plots, in which the vertical axis indicates the values for the various flow cytometric parameters, identically scaled in all the plots.

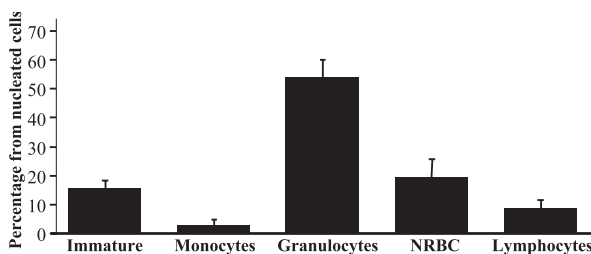


Fig 5. Cell-class: classification of normal BM samples. Ten normal BM samples, each consisting of 3×10^4 cells, were labelled and subjected to flow cytometric analysis. The cells of each sample were classified to the 23 subclasses, and accordingly into the five main classes. The bar plots show the average abundances and variations for the different main classes in 10 different BM samples.

this subclass of immature myeloid cells may represent a differentiation stage between immature myeloid cells and mature neutrophils. This was also consistent with the cluster merging dendrogram (Fig 1C), where the granulocyte subclass number 11 is relatively close to the immature subclasses.

Three subclasses of monocytes were resolved, all large cells that expressed high levels of CD45 and CD33, but varied in CD13 expression (Figs 2, 3E and 4). The monocytes subclasses comprised *c.* 3% of the BM nucleated cells (Fig 5), and their identity was confirmed by CD14 expression (data not given). We performed classification on THP-1 myelomonocytic cell line. As given in Fig 6B, >75.1% of the THP-1 cells were identified as monocytes, predominantly as subclass number 10,

which was characterized by high levels of CD13. Small amounts of the THP-1 cells (*c.* 5%) were identified as immature myeloid cells, which may correspond to the immature component of this myelomonocytic cell line (Fig 6A).

Immature cells were defined predominantly by CD34 expression. About 15% of the nucleated BM cells expressed this immature marker (Fig 5), at various levels (Fig 4), comprising seven different subclasses. We tested the classification of immature myeloid cells according to the immature subclasses. For that, K562, U937 and THP-1 cell lines, derived from myeloid leukaemia patients, were subjected to flow cytometric analysis followed by multi-dimensional classification. As given in Fig 6A, the myeloid cell lines were identified exclusively as immature myeloid subclasses. As described above, most of the monocytic THP-1 cells were identified as monocytes by our multi-dimensional analysis (Fig 6B). As expected, heterogeneity was detected among the different cell lines and within individual cell lines, as distinct subclasses were detected in all three cell lines (Fig 6A,B). In order to extend the classification of immature cells to primary immature myeloid cells, four primary AML M2 samples were included in our analysis. As given in Fig 6A, the primary AML samples were classified as immature myeloid cells, with heterogenic distribution, predominantly as subclasses number 4 and 5. Interestingly, none of the myeloid samples were classified to subclasses number 2 and 3.

Immature subclass number 3 expressed the highest CD34 levels (Figs 4 and 7), was CD45^{dim}, and had low granularity (Figs 3D and 4). Therefore its abundance was compared with the abundance of the haematopoietic progenitor cells, CD34^{bright}/SSC^{low}/CD45^{dim}, as evaluated according to the ISHAGE guidelines (Venditti *et al*, 1999) for the same BM samples. As given in Fig 7, nearly identical abundance of these cells (*c.* 1%) in normal BMs was detected both by our scheme and by routine FACS analysis according to the ISHAGE guidelines. To examine the reproducibility of our scheme in classifying cells into subclasses we have performed consecutive sample processing, acquisition and analysis of four different batches of the same BM sample. As given in Fig 7 (3*) the number of CD34⁺ progenitor cells in this subclass vary within *c.* 5%, well in line with values accepted for the CD34⁺ progenitors enumeration methods (Dzik *et al*, 1999). The CD34⁺/CD13⁺/CD33⁻ population (subclass number 4) exhibited properties previously reported for such cells, namely small cells with low granularity (Figs 3 and 4) (Gaipa *et al*, 2002).

In summary, our multi-dimensional flow cytometric analysis scheme provided high-quality classification of the major haematopoietic cell types. In addition, the analysis dissected the major classes into clear reproducible subclasses, hence pointing to differentiation and maturation processes. We have established a normal BM monocyte/myeloid template, which could be utilized to define and study deviations in relatively simple (e.g. leukaemias) or complex (e.g. myelodysplastic syndrome) haematological disorders.

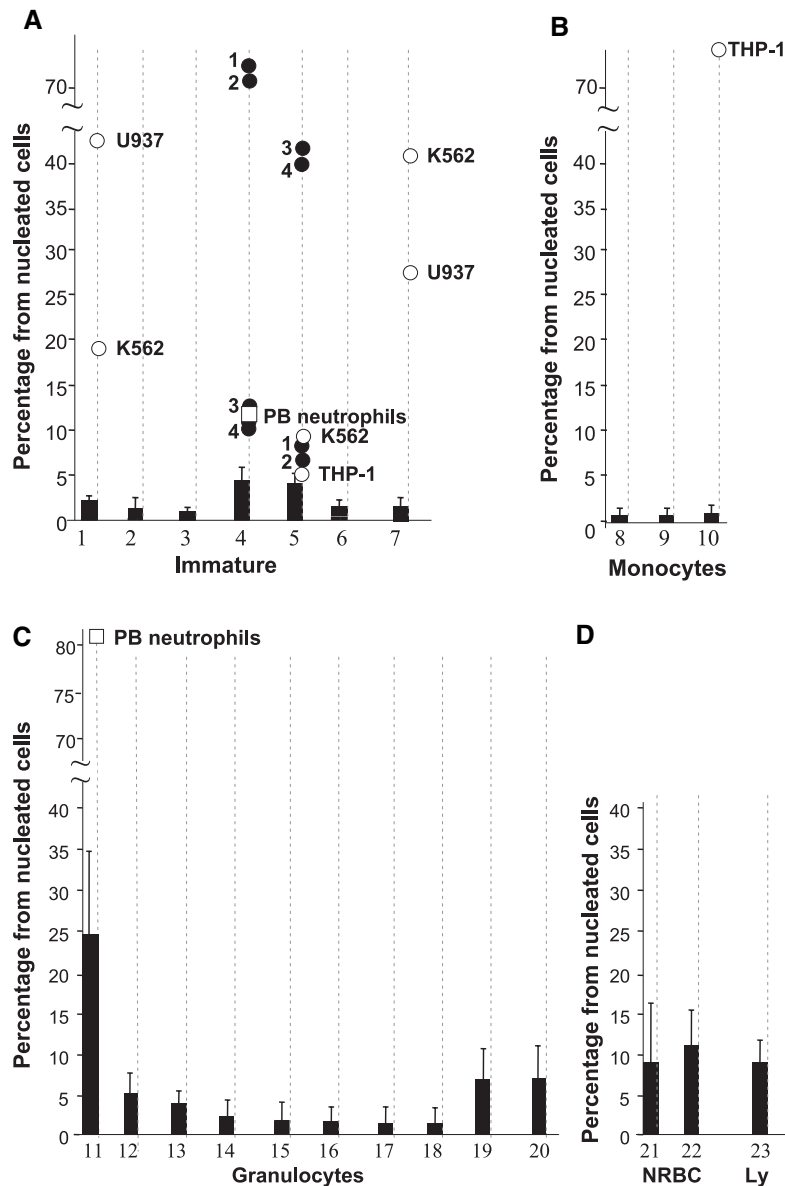


Fig 6. Cell-subclasses classification of myeloid cell lines, and normal and leukemic BM samples. Ten normal BM samples, four AML M2 primary aspirates (designated 1–4), and the myeloid cell lines K562, U937 and THP-1, each consisting of 3×10^4 cells, were labelled and subjected to flow cytometric analysis. The cells of each sample were classified into one of the 23 subclasses, including immature cells (A), monocytes (B), granulocytes (C) and NRBC and lymphocytes (D). The bar plots show the average abundances and variations for the different subclasses in the different normal BM samples. Filled circles show the subclasses abundances of the primary AML samples (designated 1–4, bold). Unfilled circles each represent a designated myeloid cell line. Note that each of the AML samples, and the myeloid cell lines, were distributed among more than one subclass. Unfilled squares show the distribution of normal peripheral blood (PB) neutrophils.

Discussion

In this study we developed and utilized a method for classification and analysis of cell types within normal BM samples. Classes and subclasses of cells were characterized in normal BM samples, and a classifier was constructed in multi-dimensional flow-cytometric parameter space. The multi-dimensional classifier was designed by considering the closeness of six-parameter values, as well as the shape of the

distribution of each subclass, resulting from intra-population variability of each parameter and correlations between parameters. We showed that the procedure yielded distributions of BM cells between five major classes, as well as classification of minor cell populations. Current enumeration schemes for $CD34^+$ progenitors are based on multi-parameter analyses performed by sequential two-dimensional gating. The common schemes, Milan-Mulhouse and ISHAGE, utilize three and seven two-dimensional gates (including isotype controls),

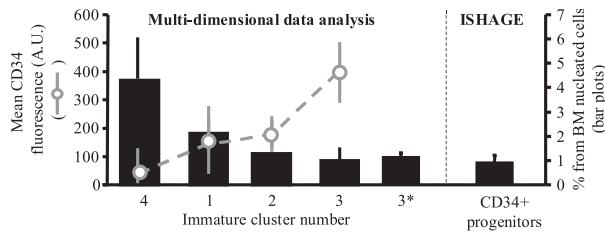


Fig 7. Analysis of immature cell populations within normal BM. Ten normal BM samples were labelled and subjected to flow cytometric analysis and cell-subclass classification. The percentages of several immature subclasses out of the total nucleated BM cells are shown (\pm SD), with the average CD34 fluorescence values (left panel). To evaluate intra-sample noise, four batches of a single normal BM sample were acquired and analysed separately. The percentages of immature subclass 3 in these four measurements from the nucleated BM cells are shown ($3^* \pm$ SD, left panel). Haematopoietic progenitors were also evaluated in the same 10 BM samples according to the ISHAGE guidelines. Percentages of these cells out of the total nucleated BM cells are shown (\pm SD) (right panel).

respectively (Venditti *et al*, 1999). Our multi-dimensional classifier allows a single step classification of CD34⁺ subpopulations, while simultaneously evaluating the whole BM pattern, including NRBC. Thus, the scheme can provide an accurate six-parameter identification of the haematopoietic progenitors, and a denominator for a dual platform enumeration scheme, namely CD45⁺ cells with the subtraction of NRBC (subclasses number 21 and 22). The method is reproducible, and may be applied to any other cell type within the BM, or any other source. While the clustering and the definition of classes and subclasses require an involvement of a knowledgeable expert, the classification process is very fast, robust and automatable. Thus, multi-dimensional classification can be applied for real-time analysis of flow-cytometry data, to easily obtain and document results that are valuable for quantitative and statistically reliable diagnostic purposes.

The majority of BM nucleated cells belong to the myeloid lineage. We classified 10 distinct subclasses of granulocytes according to their membrane marker patterns and physical properties. Most of these subclasses contain, each <5% of the total nucleated BM cells (Fig 6). Interestingly, most of the granulocytes were classified to a single subclass (number 11, Fig 6). The cells of this subclass comprise *c.* 25% of the BM nucleated cells, are relatively small, and express very high CD13, and low CD33 levels. Apparently, these cells constitute the largest granulocyte reservoir within the BM. We found that these cells were compatible with mature neutrophils. As peripheral blood neutrophils frequently contaminate BM aspirates, we attributed these cells, identified in our normal BM cohort, to both the end-point differentiation of the myeloid lineage and the presence of peripheral blood in BM aspirates. In our normal BM samples, peripheral blood contamination was relatively low (see *Materials and methods*). The complicated relationship between CD13, CD33 and CD34

expression in the myeloid subclasses, as revealed in our analysis, may point to a complex expression pattern during myeloid differentiation, rather than to the simple increase in CD33 expression followed by CD13 reduction, as depicted in previous studies and text books (Hoffman *et al*, 2003). The small CD34⁺/CD13⁺/CD33⁻ population may indicate that myeloid progenitors show initially elevated CD13 with expansion, while the levels of CD34 drop. Future studies based on preparative FACS will enable morphological and molecular characteristics to be linked with the granulocyte subclasses. Once established, such a flow cytometric-based analysis scheme should result in a rapid and reproducible analysis of normal granulocyte differentiation and maturation.

The multi-dimensional flow cytometric analysis scheme was able to classify accurately the minor monocytes population within normal BM samples, in a reproducible manner between normal individuals. This classification was confirmed by accurate identification of the THP-1 monocytic cell line. Hierarchical clustering merged the immature subclass 7 with the monocytic lineage (Fig 1C). Based on multi-dimensional signature similarities, these cells, which comprise *c.* 2% of the BM nucleated cells, may correspond to the common origin of myeloid/monocytic lineage. These cells are large, have low granularity, and express relatively high levels of CD45 with CD13 and CD33. Three distinct subclasses of monocytes were identified, having a significant difference with respect to CD13 expression levels. CD13 has been assigned to the function of mature monocytes, and associated with monocytes differentiation (Riemann *et al*, 1999). These subclasses may represent normal differentiation pattern of monocytes within the BM, ending with a population of high CD13 expressing monocytes. While previous studies indicated that CD13 is regulated predominantly at the RNA level within the BM (Dybkaer *et al*, 2001), we clearly showed that several distinct subclasses of myeloid cells and monocytes express various levels of surface CD13.

Flow cytometry is widely used, both in routine medical laboratory workouts and as a prime scientific tool. However, clinical guidelines that outline algorithms for the diagnosis and classification of haematopoietic malignancies do not include flow cytometry as a mandatory diagnostic method, but rather as an accessory tool (Appelbaum *et al*, 2000; Vardiman *et al*, 2002). Cell classification schemes (e.g. World Health Organization, French-American-British, National Comprehensive Cancer Network) are based primarily on morphological evaluation, cytogenetics and histochemistry (Appelbaum *et al*, 2000; Vardiman *et al*, 2002). Several reasons may account for the limited application of flow cytometry for the diagnosis and prognosis of haematological disorders. First, the current definition of cell populations in FACS histograms is subjective. Flow cytometric application in disease prognosis is limited to studies based on single antibody analyses such as CD34 enumeration or ZAP-70 expression in chronic lymphocytic leukaemia (Crespo *et al*, 2003; Orchard *et al*, 2004). Even in the case of a single label, conflicting definitions of a positive

population, established either by a reference cell population (Crespo *et al*, 2003) or by isotype control (Orchard *et al*, 2004), may hinder uniform prognostic evaluation. Second, clinical cytometry forums recommend the use of 20–24 different antibodies to obtain a reliable diagnosis of acute leukaemias (Braylan *et al*, 2001). The need for such a high number of markers for accurate cell identification is because of the small number of cell types that can be identified by standard flow cytometry analysis methods. Consequently, this probably hindered the establishment of uniform flow cytometry schemes in research and clinical laboratories. Therefore, a link between antigen expression patterns and the clinical course of haematological disorders (e.g. AML), specifically prognosis, cannot be established today. As shown here, the multi-dimensional clustering approach can reduce significantly the number of markers required for the identification of cell types/lineages within the BM, turning the development of uniform clinical flow cytometry schemes feasible. Such schemes will extend the single parameter values, namely percentage of positive cells and mean fluorescence intensity, into six-dimensional ‘signatures’ of defined populations (either normal or abnormal). This, in turn, may also yield clinical information, based on large sample cohorts, leading to prognostic significance of the data.

In addition to comprehensive classification and quantitative assessments of the number of cells in each subclass, the multi-dimensional parametric space enables the definition of specific cell type signatures for abnormal populations of individual patients diagnosed with a haematological disorder such as leukaemia. This signature can be utilized to monitor treatment of the disease, and to follow with high specificity and sensitivity minor residual populations during recession stages. The variability between leukaemias, and the need for identifying recurrence because of proliferation of residual cells, as opposed to newly mutated clones, can be addressed by such multi-parametric signatures.

Multi-parametric flow cytometry, when combined with multi-dimensional data analysis, can provide an accurate identification of minor cell populations within the BM (e.g. CD34⁺ progenitors, monocyte subclasses). Moreover, it can incorporate the quantitative analysis of such minor (but important) cell populations, within a comprehensive pattern analysis of the whole haematopoietic BM reservoir. Hence, quantitative assessment of differentiation, maturation and expansion processes, that requires simultaneous identification and measurement of multiple cellular populations in variable sample cohorts, is within reach.

Acknowledgements

This study was partly supported by the Yad Avraham Centre for Cancer Diagnostics and Therapy. BG holds the E. Netter Chair in Cell and Tumour Biology. ZK is the Israel Pollak Professor of Biophysics.

This work is dedicated to the memory of Yael (Zamir) Cohen.

Supplementary material

The following material is available from <http://www.blackwellpublishing.com/products/journals/suppmat/bjh/bjh5471/bjh5471sm.htm>

Appendix S1. A practical protocol for multi-dimensional analysis of bone marrow cells

Appendix S2. Programs documentation

References

- Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. & Levine, A.J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 6745–6750.
- Appelbaum, F.R., Baer, M.R., Carabasi, M.H., Coutre, S.E., Erba, H.P., Estey, E., Glenn, M.J., Kraut, E.H., Maslak, P., Millenson, M., Miller, C.B., Saba, H.I., Stone, R. & Tallman, M.S. (2000) NCCN Practice Guidelines for Acute Myelogenous Leukemia. *Oncology*, **14**, 53–61.
- Bain, B.J. (1996) The bone marrow aspirate of healthy subjects. *British Journal of Haematology*, **94**, 206–209.
- Batinic, D., Marusic, M., Pavletic, Z., Bogdanic, V., Uzarevic, B., Nemet, D. & Labar, B. (1990) Relationship between differing volumes of bone marrow aspirates and their cellular composition. *Bone Marrow Transplantation*, **6**, 103–107.
- Beckman, R.J., Salzman, G.C. & Stewart, C.C. (1995) Classification and regression trees for bone marrow immunophenotyping. *Cytometry*, **20**, 210–217.
- Blatt, M., Wiseman, S. & Domany, E. (1996) Superparamagnetic clustering of data. *Physical Review Letters*, **76**, 3251–3254.
- Boddy, L., Wilkins, M.F. & Morris, C.W. (2001) Pattern recognition in flow cytometry. *Cytometry*, **44**, 195–209.
- Borowitz, M.J., Bray, R., Gascoyne, R., Melnick, S., Parker, J.W., Picker, L. & Stetler-Stevenson, M. (1997) U.S.–Canadian Consensus recommendations on the immunophenotypic analysis of hematologic neoplasia by flow cytometry: data analysis and interpretation. *Cytometry*, **30**, 236–244.
- Braylan, R.C., Orfao, A., Borowitz, M.J. & Davis, B.H. (2001) Optimal number of reagents required to evaluate hematology neoplasias: results of an international consensus meeting. *Cytometry*, **46**, 23–27.
- Brown, G., Capellaro, D. & Greaves, M. (1975) Leukemia-associated antigens in man. *Journal of the National Cancer Institute*, **55**, 1281–1289.
- Bullinger, L., Dohner, K., Bair, E., Frohling, S., Schlenk, R.F., Tibshirani, R., Dohner, H. & Pollack, J.R. (2004) Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *New England Journal of Medicine*, **350**, 1605–1616.
- Cantor, H., Simpson, E., Sato, V.L., Fathman, C.G. & Herzenberg, L.A. (1975) Characterization of subpopulations of T lymphocytes. I. Separation and functional studies of peripheral T-cells binding different amounts of fluorescent anti-Thy 1.2 (theta) antibody using a

- fluorescence-activated cell sorter (FACS). *Cellular Immunology*, **15**, 180–196.
- Crespo, M., Bosch, F., Villamor, N., Bellosillo, B., Colomer, D., Rozman, M., Marce, S., Lopez-Guillermo, A., Campo, E. & Montserrat, E. (2003) ZAP-70 expression as a surrogate for immunoglobulin-variable-region mutations in chronic lymphocytic leukemia. *New England Journal of Medicine*, **348**, 1764–1775.
- De Rosa, S.C. & Roederer, M. (2001) Eleven-color flow cytometry. A powerful tool for elucidation of the complex immune system. *Clinics in Laboratory Medicine*, **21**, 697–712.
- De Rosa, S.C., Herzenberg, L.A. & Roederer, M. (2001) 11-color, 13-parameter flow cytometry: identification of human naive T cells by phenotype, function, and T-cell receptor diversity. *Nature Medicine*, **7**, 245–248.
- De Rosa, S.C., Brenchley, J.M. & Roederer, M. (2003) Beyond six colors: a new era in flow cytometry. *Nature Medicine*, **9**, 112–117.
- Dybkaer, K., Olesen, G., Pedersen, F.S. & Kristensen, J.S. (2001) Stromal-mediated down-regulation of CD13 in bone marrow cells originating from acute myeloid leukemia patients. *European Journal of Haematology*, **66**, 168–177.
- Dzik, W., Sniecinski, I. & Fischer, J. (1999) Toward standardization of CD34⁺ cell enumeration: an international study. Biomedical Excellence for Safer Transfusion Working Party. *Transfusion*, **39**, 856–863.
- Engelhardt, M., Lubbert, M. & Guo, Y. (2002) CD34(+) or CD34(-): which is the more primitive? *Leukemia*, **16**, 1603–1608.
- Gaipa, G., Coustan-Smith, E., Todisco, E., Maglia, O., Biondi, A. & Campana, D. (2002) Characterization of CD34⁺, CD13⁺, CD33⁻ cells, a rare subset of immature human hematopoietic cells. *Haematologica*, **87**, 347–356.
- Hoffman, R., Benz, E.J., Shatill, S.J., Furie, B., Harvey, J.C., Silberstein, L.E. & McGlave, P. (eds) (2003) *Hematology: Basic Principles and Practice*. Churchill Livingstone, Philadelphia, PA.
- Hrusak, O. & Porwit-MacDonald, A. (2002) Antigen expression patterns reflecting genotype of acute leukemias. *Leukemia*, **16**, 1233–1258.
- Jennings, C.D. & Foon, K.A. (1997) Recent advances in flow cytometry: application to the diagnosis of hematologic malignancy. *Blood*, **90**, 2863–2892.
- Mingari, M.C., Vitale, C., Romagnani, C., Falco, M. & Moretta, L. (2001) p75/AIRM1 and CD33, two sialoadhesin receptors that regulate the proliferation or the survival of normal and leukemic myeloid cells. *Immunological Reviews*, **181**, 260–268.
- Orchard, J.A., Ibbotson, R.E., Davis, Z., Wiestner, A., Rosenwald, A., Thomas, P.W., Hamblin, T.J., Staudt, L.M. & Oscier, D.G. (2004) ZAP-70 expression and prognosis in chronic lymphocytic leukemia. *Lancet*, **363**, 105–111.
- Penninger, J.M., Irie-Sasaki, J., Sasaki, T. & Oliveira-dos-Santos, A.J. (2001) CD45: new jobs for an old acquaintance. *Nature Immunology*, **2**, 389–396.
- Richard, O., Duda, D., Stork, G. & Peter, E.H. (eds) (1999) *Pattern Classification*, 2nd edn. Wiley, John & Sons, Inc., New York, NY.
- Riemann, D., Kehlen, A. & Langner, J. (1999) CD13—not just a marker in leukemia typing. *Immunology Today*, **20**, 83–88.
- Valet, G.K. & Hoffkes, H.G. (1997) Automated classification of patients with chronic lymphocytic leukemia and immunocytoma from flow cytometric three-color immunophenotypes. *Cytometry*, **30**, 275–288.
- Valk, P.J., Verhaak, R.G., Beijnen, M.A., Erpelinck, C.A., Barjesteh van Waalwijk van Doorn-Khosrovani, S., Boer, J.M., Beverloo, H.B., Moorhouse, M.J., van der Spek, P.J., Lowenberg, B. & Delwel, R. (2004) Prognostically useful gene-expression profiles in acute myeloid leukemia. *New England Journal of Medicine*, **350**, 1617–1628.
- Vardiman, J.W., Harris, N.L. & Brunning, R.D. (2002) The World Health Organization (WHO) classification of the myeloid neoplasms. *Blood*, **100**, 2292–2302.
- Venditti, A., Battaglia, A., Del Poeta, G., Buccisano, F., Maurillo, L., Tamburini, A., Del Moro, B., Epiceno, A.M., Martiradonna, M., Caravita, T., Santinelli, S., Adorno, G., Picardi, A., Zinno, F., Lanti, A., Bruno, A., Suppo, G., Franchi, A., Franconi, G. & Amadori, S. (1999) Enumeration of CD34⁺ hematopoietic progenitor cells for clinical transplantation: comparison of three different methods. *Bone Marrow Transplantation*, **24**, 1019–1027.
- Weir, E.G. & Borowitz, M.J. (2001) Flow cytometry in the diagnosis of acute leukemia. *Seminars in Hematology*, **38**, 124–138.
- Wells, D.A., Benesch, M., Loken, M.R., Vallejo, C., Myerson, D., Leisenring, W.M. & Deeg, H.J. (2003) Myeloid and monocytic dyspoiesis as determined by flow cytometric scoring in myelodysplastic syndrome correlates with the IPSS and with outcome after hematopoietic stem cell transplantation. *Blood*, **102**, 394–403.