CYTOMETRY Journal of Quantitative Cell Science PART A

Challenges in the Multivariate Analysis of Mass Cytometry Data: The Effect of Randomization

Georgios Papoutsoglou,[1*] Vincenzo Lagani,[2,3] Angelika Schmidt,[4†] Konstantinos Tsirlis,[5] David-Gómez Cabrero,[4,6] Jesper Tegnér,[4,7] Ioannis Tsamardinos[1,3]

[1]Computer Science Department, University of Crete, Heraklion, Greece

[2]Institute of Chemical Biology, Ilia State University, Tbilisi, Georgia

[3]Gnosis Data Analysis PC, Heraklion, Greece

[4]Unit of Computational Medicine, Center for Molecular Medicine, Department of Medicine Solna, Karolinska Institutet, Karolinska University Hospital and Science for Life Laboratory, Stockholm, Sweden

[5]Department of Computer Science, University College London, London, UK

[6]Translational Bioinformatics Unit, Navarrabiomed, Pamplona, Spain

[7]Biological and Environmental Sciences and Engineering Division, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

Received 20 March 2019; Revised 27 August 2019; Accepted 18 September 2019

Grant sponsor: FP7 Ideas: European Research Council, Grant number: 617393

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Georgios Papoutsoglou, University of Crete – Department of Computer Science, Voutes Campus, GR-70013 Heraklion, Crete, Greece Email: papoutsoglou@csd.uoc.gr

†Present address of Angelika Schmidt: Institute for Immunology, Biomedical

Abstract

Cytometry by time-of-flight (CyTOF) has emerged as a high-throughput single cell technology able to provide large samples of protein readouts. Already, there exists a large pool of advanced high-dimensional analysis algorithms that explore the observed heterogeneous distributions making intriguing biological inferences. A fact largely overlooked by these methods, however, is the effect of the established data preprocessing pipeline to the distributions of the measured quantities. In this article, we focus on randomization, a transformation used for improving data visualization, which can negatively affect multivariate data analysis methods such as dimensionality reduction, clustering, and network reconstruction algorithms. Our results indicate that randomization should be used only for visualization purposes, but not in conjunction with high-dimensional analytical tools.

Key terms

mass cytometry; pre-processing; high dimensional data analysis; randomization; dimensionality reduction; clustering; network reconstruction; method development

Single cell cytometry allows the detection of cell components in a high-throughput fashion. One of its latest versions is Cytometry by Time Of Flight (CyTOF) (1). The advantage of CyTOF compared to traditional flow cytometry is that high atomic weight metal reporters typically not found in a biological sample are employed for cell tagging, allowing the quantification of more than 40 cell parameters simultaneously. Such large number of parameters enables this technology to provide multivariate data sets with emerging properties that are well suited to advanced computational analysis (2,3). For example, unsupervised learning techniques like clustering and dimensionality reduction are typically used for cell phenotyping (4,5). In combination with statistical tests or supervised learning approaches, these methods are also employed for associating phenotypes or clinical outcomes to relevant cell subsets or protein markers (6,7). Clustering and dimensionality reduction are also commonly employed to visualize patterns in the data, marker relationships in the high-dimensional space or the phenotypic progression trajectories of cell subsets (8). Very recently, network–based methods have also been applied on CyTOF data, for automatic cell population identification (9) and the prediction of protein signaling networks using automated causal discovery algorithms (10).

Despite the accelerated development of CyTOF-dedicated analysis methods, there is still no well-established data preprocessing consensus. This is mainly because the preceding standardization of experimental procedures is still in its infancy (11,12). In general, there are at least three distinct sources of technical variation in CyTOF. The first is the drop in the instrument sensitivity and the change in oxidation rate over long sample running times that causes signal fluctuations (13). Second,

Center, Ludwig-Maximilians-Universität München, Planegg-Martinsried, Germany.

Published online 5 November 2019 in Wiley Online Library (wileyonlinelibrary.com)

DOI: 10.1002/cyto.a.23908

the instrument sensitivity across the mass range of measurement that also varies between different instruments (14). Third, is the interference artifacts (spillovers) between mass detection channels (15). Even though these sources of variation have been described, most of the developed preprocessing methods are still immature. On the other hand, there are some tasks that have been adapted from flow cytometry data analysis workflows. One such task is data randomization whose purpose is to transform discrete measurement values into continuous ones in order to aid the visualization of bivariate distributions. These plots are typically used for the manual assignment of cells to cell types—better known as gating (16).

Any systematic effect of CyTOF preprocessing tasks to the distributions of the measured quantities has been largely overlooked, particularly randomization. This is also evident from the fact that in the majority of publications, the randomization settings are nowhere reported. As a result, the faithful reproduction of the true distributions in public data is problematic. In turn, the proper functioning of high dimensional analysis methods developed on the basis of randomized data is also questionable, because their consistency rests in the assumptions made about the underlying data distributions.

In this article we test and discuss the effects of usual CyTOF data preprocessing strategies to multivariate analyses. We focus on randomization, the only routinely used preprocessing task which apparently has no theoretical or practical justification for computational analysis. We employ in-house generated CyTOF data and compare results from several multivariate analyses, namely, dimensionality reduction, clustering, and network reconstruction (17), before and after applying randomization. Our results show that randomization heavily influences the results of multivariate analytical methods. Furthermore, publicly available CyTOF data sets appear to typically provide solely randomized data, which cannot be easily reverted to a non-randomized state.

Materials and Methods

CyTOF Data Preprocessing: From Raw Data to Data Ready for Analysis

Generating data regarding each cell is a quite complex process in CyTOF that involves raw measurements' calibration, noise subtraction, and cell detection tasks (18). Figure 1A,B give an outline of this process. In turn, Figure 1C–D show the succeeding preprocessing tasks arranged in a typical order for data bivariate visualization and gating. In more detail, raw (.FCS) CyTOF data initially consist of count values that indicate the abundance of each measured protein per identified cell. Then, because of the signal decay over time, the first preprocessing task one needs to perform is data normalization (Fig. 1C). Finck et al. have shown that this is possible by injecting control beads to each biological sample (13). On this basis, there are currently two normalization approaches: the method presented in Finck et al. and the Fluidigm normalizer built into the machine dedicated software. Both approaches assume that the decline in machine sensitivity is a linear function of acquisition time and, so, multiplying the data in each

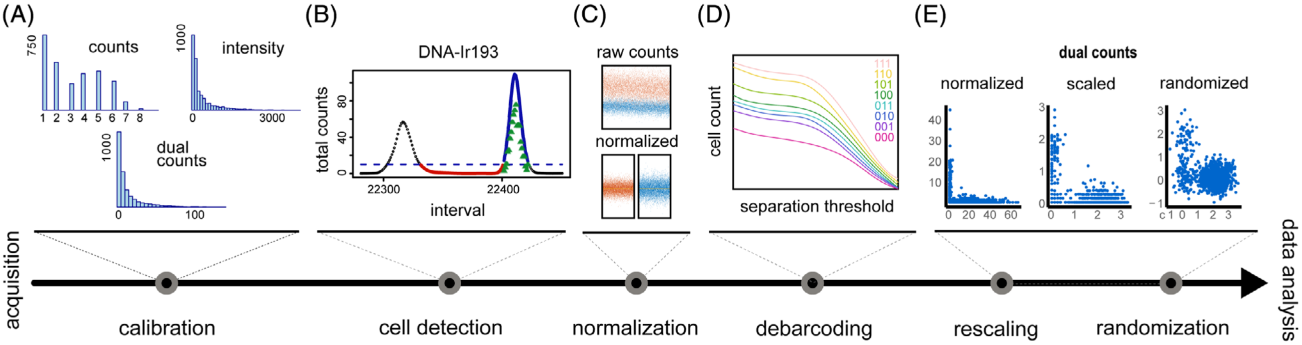

Figure 1. Preprocessing roadmap for data visualization and gating. (A) In dual count calibration the intensity values are used to predict the true ion counts. (B) To identify cells (called cell events), the counts are summed and smoothed. A cell event (blue curve) is found when the run of smoothed counts is larger than the noise threshold (dashed line) for more than 10 (by default) consecutive time intervals. The dual counts before a cell event (red curve) are regarded as background noise and used for (optional) noise correction. The green triangles indicate the abundance of a marker in that specific cell event. (C) Normalization corrects for signal decay over large machine running times. (D) De-barcoding de-convolves the multiplexed experiments according to a binary barcoding scheme. (E) Data rescaling is performed using the arcsinh transform to bring abundance distributions to comparable ranges and reveal patterns in the data. Randomization is applied to smoothe striation patterns in low abundance distributions. [Color figure can be viewed at wileyonlinelibrary.com]

cell by a time-dependent global standard will be correcting signal fluctuations. Their difference lays in that the former employs the median bead counts as global standard calculated across experimental files to perform separate or batch sample normalization while the latter employs reference values unique to each lot of beads, as determined by the device manufacturer, and can only be applied on a per sample basis.

A second task is cell de-multiplexing and filtering if, for improving the quality of the data, several experiments are multiplexed in one sample tube (Fig. 1D) (19). This task, also known as debarcoding, is necessary not only because cells need to be assigned back to the experimental samples they correspond to, but also because contaminating cross-sample information such as low signal debris and cell doublets need to be filtered out.

The third task is data rescaling transformation to allow their adequate representation across the full range of abundance values (Fig. 1E). In general, protein markers tend to have strongly skewed distributions with varying ranges of abundance. Therefore, the inverse hyperbolic sine (arcsinh), adjusted by a co-factor of five, is most frequently used because it serves as a linear transform to abundances in the low range and as a log transform in the high range, with a smooth transition between them (20).

Fourth is randomization that it is performed to avoid the poor display of low abundance distributions (Fig. 1E). Essentially, CyTOF data contain high numbers of zero and close to zero counts. As a result, bivariate plots show large numbers of points piling-up around integer values creating striation patterns. In turn, cell populations become difficult to delineate during manual gating. The purpose of randomization, then, is to spread the abundance values so that data resolution are locally enhanced. There are three types of randomization. Type 1, that is to spread every count number x evenly in the interval (x-1,x) by use of a uniform distribution. Type 2, that is to substitute every value x with a random number drawn from a Gaussian density with mean x and user-defined standard deviation σ. Type 3, that is to spread every zero count by use of a random number generated using the negative absolute value of the Gaussian with mean zero and user-defined variance σ that is, abs($N[0,\sigma]$). Currently, the device software (version 6.7.1014) offers applying Type 1 randomization right after cell detection (Fig. 1B) and either Type 1 or Type 2, and Type 3 separately after normalization (Fig. 1C). Its default setting is to apply Type 1 randomization in both cases.

There are also some tasks that we did not mention in detail. These are: (1) the preprocessing tasks for generating the data regarding each cell from raw measurements, namely, dual count calibration, noise subtraction and cell detection (Fig. 1A, B); (2) the detection of dead cells and cell doublets or debris, which are filtered out through manual gating, and the detection of data discontinuities due to discontinuous sample injections to the machine, which can be resolved by data concatenation; and (3) two important methods for batch effect removal (21) and spillover correction (22) that may in the future become commonly employed but have not reached that point yet.

CyTOF Data Type: The Effect of Randomization

From these pre-processing tasks, randomization is the only one not justified by any need of correcting possible biases or providing a more precise quantification. It is rather a heuristic regularly used to assist cytometry experts in improving bivariate visualizations during gating. This task, however, considerably changes the underlying data distributions. Initially, the raw data are count data. Normalizing the counts subtly changes their form from count to discrete since each count value is multiplied by a continuous correction factor. Demultiplexing and arcsinh transformation preserve this form as they act only as a filtering and rescaling function, respectively. Randomization, however, dramatically changes the data because it injects unnecessary noise, essentially transforming the data from discrete to continuous. How this change affects the output of downstream analyses has not yet been resolved and a systematic investigation is required.

High Dimensional Analyses

We selected a set of multivariate computing algorithms widely used for dimensionality reduction, clustering and network learning. Specifically, we exemplify the effect of randomization on linear and nonlinear dimensionality reduction by principal components analysis (PCA), t-stochastic neighborhood embedding (tSNE) analysis (23), an extension of PCA known as joint and individual variation explained (JIVE) (24) and multidimensional scaling (MDS). The first two methods are typically used for visualizing high dimensional phenotypes in two or three dimensions. PCA distills the high dimensional space by linearly transforming the data to a new coordinate system whose coordinates explain the variance of the data in descending order. tSNE seeks a low dimensional representation of the data that best preserves their geometry in the high dimensional space. On the other hand, JIVE is capable of identifying the amount of joint and individual structure characterizing two data matrices. Finally, MDS is used for measuring the level of similarity between different data sets.

Since clustering is very important in automated cell population identification we also tested the effect of randomization on the robustness of the deterministic spanning-tree progression analysis for density-normalized events (SPADE) algorithm and a meta-clustering approach (5,25). SPADE is the best-known clustering approach in CyTOF while the meta-approach was best performing in a recent review (5). Both algorithms perform an over-clustering step. Their internal mechanics, however, differ. In brief, SPADE initially performs a density-based downsampling such that the generated subset of data potentially encapsulates all possible cell phenotypes. Then, it groups this subset into cell types by applying k-means or hierarchical agglomerative clustering. The generated clusters are finally linked and visualized by use of a minimum-spanning tree. The meta-clustering, on the other hand, is a two-stage process by which the data are mapped onto a large grid of points using self-organizing maps (SOM), first. Because this mapping displays emergent properties, one can perform further clustering operations. Therefore, meta-clustering is performed that groups the points in the grid and leads to a reduced latent feature space.

Lastly, we evaluate the effect of randomization on network learning by employing three reconstruction methods, namely, relevance networks (RNs), the graphical Lasso (GLasso) method, and the really fast causal inference (RFCI) algorithm (26–28). All methods accept as input a matrix whose rows refer to sample data (i.e., cells) and columns to measured quantities (i.e., markers), and provide as output a graph whose nodes represent the cell markers. Their inner operation, however, and network semantics largely differ. RNs simply quantify the mutual association between each pair of proteins. If sufficiently strong, an undirected edge connects the pair in the resulting graph. GLasso attempts to designate the partial correlations between pairs of markers that are significantly different from zero when conditioned upon all others. The result is also an undirected graph whose edges indicate that the association between the nodes cannot be explained by other variables. Lastly, RFCI is a constraint-based, causal discovery method. Initially, it builds an undirected graph whose edges define pairs of variables that remain associated after conditioning on any other possible subset. Then, a set of rules is used to mark the endpoints of each edge based on the possible causal nature of the relationship, that is, presence of a causal effect, attributed to unmeasured latent quantities or unresolved (29). The resulting graph is known as maximal ancestral graph (MAG); interested readers are referred to (30) for further information.

Evaluation Parameters

In our evaluations we used the same set of data on peripheral blood mononuclear cells (PBMCs) obtained from a sample of one healthy donor, before and after randomization had been applied. Data were preprocessed having the automatic randomization disabled. For data normalization we used the algorithm of Finck et al. whereas to remove cell doublets and dead cells, the data were manually pre-gated (Supporting Information Fig. S13). Both Type 1 and Type 2 randomization were applied on the data. For Type 2 randomization we substituted every value x with a random number drawn from a Gaussian density centered at x and a standard deviation of 1, which is the default in the device software. Then, each multivariate analysis algorithm was applied in turn on both randomized and non-randomized data using its default input parameter values. Prior to each analysis the data were transformed using the hyperbolic arcsin with a cofactor of five. To investigate the reasonable limits of the effect of randomization we also employ data preprocessed for bivariate visualization and gating as shown in Figure 1. For this, we applied Type 2 randomization with a standard deviation of 0.3. We will refer to this randomization scheme as the maximal randomization type.

Finally, to ensure reproducibility and comparability of results no downsampling of cells was performed. In addition, for the algorithms that are sensitive to random starts like tSNE or meta-clustering, we defined the same random seed before each run. Supporting Information Table S1 lists the availability of software implementations; for more details, the interested reader is referred to the corresponding manuscripts. Supporting Information Table S2 indicates the markers used in each multivariate analysis.

Evaluation Metrics

The agreement between dimensionality reduction results was determined by visual inspection of the respective low dimensional geometries. For clustering results we applied each algorithm once on each data set and used the adjusted rand index (ARI) and the F_1 measure. The ARI is the adjusted-for-chance form of the RAND Index (31). The Rand index, itself, specifies the probability of agreement between two partitions and is defined as the percentage of pairwise assignments that are true (positive or negative). Because the chance of random agreement can be high, the baseline can be nonzero. To establish a proper baseline an adjustment-for-chance is possible assuming the generalized hypergeometric distribution as the null model. After correction ARI values range between −1 and 1 where, 1 indicates total agreement between two clusters and zero or less than zero indicates that the agreement is equal or less than what is expected if the two clusters were drawn at random. The F_1 measure, on the other hand, quantifies the overlap between two subsets. It is defined as the weighted harmonic mean of precision and recall for a single cluster. Here, precision measures the proportion of cells in a randomized cluster that are comprised of cells from a non-randomized cluster. Respectively, recall measures the proportion of cells in the non-randomized cluster that were found in the randomized cluster. The F_1 measure ranges from 0 to 1 where, 1 indicates that the assignment of cells to a given cluster is exactly the same with no false positive or false negative events. To match the original and randomized clusters we used the Hungarian algorithm on the calculated F_1 measures as in (32).

To evaluate the effect on network reconstruction results we apply each algorithm in turn on all data sets and compare the obtained networks in terms of number of different edges or structural hamming distance (SHD) (33,34). The first metric is suitable for algorithms that output undirected graphs i.e. RNs and GLasso, while the SHD is devised for comparing causal networks, by counting the number of modifications (removing or adding edges, changing endpoints) needed in order to transit from one network to the other. The SHD reduces to the number of different edges for undirected graphs.

Data set

Ethics statement

Anonymized healthy donor buffy coats were purchased from the Karolinska University Hospital (Karolinska Universitetssjukhuset, Huddinge), Sweden. Research was performed according to the national Swedish ethical regulations (ethical review act, SFS no. 2003:460).

Cell isolation, culture, and storage

PBMCs were isolated from buffy coats by Ficoll-based density gradient centrifugation according to standard procedures, followed by monocyte depletion through plastic adherence and platelet depletion by low speed centrifugation. PBMCs were then cultured in RPMI medium containing 10% FCS at 37°C/5% CO_2 and to boost intracellular cytokine expression, PBMCs were stimulated for 4.25 h with Phorbol 12-myristate 13-acetate (10 ng/ml; Sigma-Aldrich, St. Louis, MO) and ionomycin (375 ng/ml; Sigma Aldrich) in the presence of the protein transport inhibitor Brefeldin A (1x BD GolgiPlug, BD Biosciences). Cells were harvested by centrifugation, washed with PBS, and fixed using PBMC fixation/wash buffer set (CytoDelics AB) according to the manufacturer's recommendations. Subsequently, cells were washed in FCS, resuspended in freezing medium (10% DMSO, 90% FCS) and cryopreserved until use.

Sample preparation for mass cytometry

Samples were prepared according to standard procedures by the SciLifeLab National Mass Cytometry facility in Stockholm (Sweden, http://cytof.scilifelab.se/). In brief, cells were thawed using an automated system (Biocision) in RPMI medium supplemented with FCS, penicillin–streptomycin and benzonase (Sigma-Aldrich), following which cells were resuspended in CyFACS buffer (PBS with 0.1% BSA, 0.05% sodium azide and 2 mM EDTA). After thawing, samples underwent barcoding, Fc receptor blocking, surface marker staining, permeabilization, intracellular cytokine, and DNA staining, and addition of EQ Four Element Calibration Beads (Fluidigm) as described in (35). Cells were acquired on a CyTOF2 (Fluidigm) mass cytometer, CyTOF software version 6.0.626 with noise reduction, a lower convolution threshold of 200, event length limits of 10–150 pushes, a sigma value of 3, and flow rate of 0.045 ml/min.

Mass cytometry antibodies

Purified antibodies were obtained as carrier protein-free formulations and then coupled to lanthanide metals using the MaxPar X8 antibody conjugation kit (Fluidigm Inc.) according to manufacturer's recommendations. Alternatively, metal-conjugated antibodies were purchased from Fluidigm. The antibodies used for this study are listed in Supporting Information Table S2.

Results

Randomization Is Commonly Used in Public Datasets but Not in a Principled Manner

To show the widespread use of randomization, we surveyed the 100 most cited works in the field that employed or developed multivariate algorithms and recovered all 15 of them whose data are available online (see Supporting Information Table S3 for respective citations). Supporting Information Figure S1 displays visual proof of randomization in these data sets, signifying how widespread randomization is in public data. Most notably, none of the 15 publications explicitly stated that the provided data underwent randomization. To the best of our knowledge, the only non-randomized public data set has been announced in (36), during the preparations of this article.

Supporting Information Figure S1 also shows that there is no consensus as to how and at which stage randomization is applied. For example, although Type 1 randomization seems to have been used in all data sets, in some of them randomization may have been applied twice; probably before and after normalization. To illustrate the lack of consensus further, we plotted the density of abundance values around the origin for one marker from each of these data sets. Supporting Information Figure S2 clearly shows the large variability in the application of data randomization by the community.

Randomization Effect on Data Dimensionality Reduction

For assessing the effect of randomization on dimensionality reduction, we first performed PCA. Figure 2A illustrates the cell densities in the first two principal components when the original and randomized data are used. Any difference is almost indistinct. However, the variance that the first two principal components explain is low and no more than 40% in all cases (Supporting Information Fig. S3). For this reason, we explored for linear decomposition differences by jointly analyzing the non-randomized and each randomized data sets using JIVE. Figure 2B depicts the aggregate variance decomposition of the non-randomized and randomized data. The joint structure is only 60–65% in all cases. As expected, the noise injected on the data by randomization causes the residual structure to explain between 27 and 32% of the variance in the following sequence: Type 1 (27,6%), Type 2 (28.6%) and maximal randomization (31.1%). At the same time, about 10% was attributed to the residual structure without randomization in all cases.

Figure 2C illustrates the similarity between the non-randomized and randomized data sets based on the MDS of the median marker abundance values. As expected, the data sets where Type 2 randomization was employed are closely placed. The first dimension (MDS1) separates well the two randomization types indicating that the effect of randomization is expected to be similar between the Type 2 and the maximal randomization and different between the latter two and Type 1 randomization. On the other hand, the second dimension (MDS2) separates the non-randomized from the randomized data signifying the existence of a randomization effect as in Figure 2B.

To further investigate the effect on dimensionality reduction we performed a nonlinear analysis using tSNE. Figure 2D shows the cell density in the low dimensional embeddings generated by the non-randomized and randomized data. Although the embedding geometries are similar, several differences can be observed. Most importantly, the contour lines of the non-randomized embedding indicate numerous local maxima implying that it may be capturing more cell subsets than the randomized ones. To get a better view of this, we overlaid the density of each marker on the respective embedding geometries. All markers analyzed are shown in Supporting Information Figure S4 wherein the distribution of protein abundances between all cases are

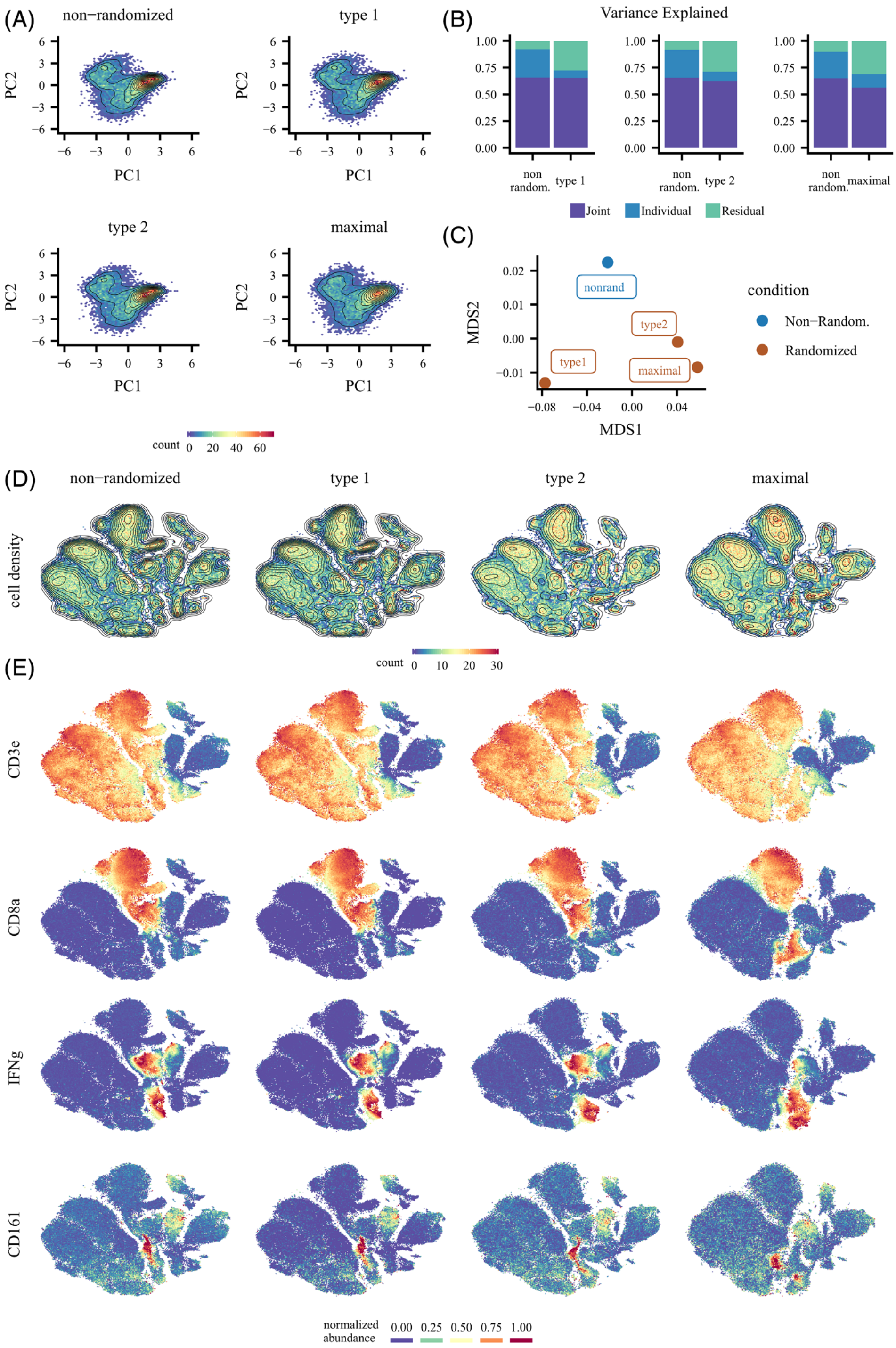

Figure 2. Effect of randomization on dimensionality reduction. (A, B) Effect on linear dimensionality reduction. (A) Depicts the cell density in the first two principal components when the non-randomized and randomized data are employed. Color-coding and contour lines denote any difference when the two data sets are independently decomposed. The result from the joint decomposition is shown in (B), where, the percentages of the variability attributable to joint, individual, and residual structure is illustrated in purple, blue, and green, respectively. (C) Multidimensional scaling plot showing the similarities between the nonrandomized and randomized data sets. (D, E) effect on tSNE analysis. (D) Depicts the cell densities when the original and randomized variable space is reduced down to two dimensions. Color-coding and contour lines capture the local differences between the embeddings. For example, the more the number of peaks in the embedding, the more cell populations it may be capturing. The columns in (E) depict the spatial distribution of four different-sized groups on each low dimensional geometry. In particular, the color-coded abundance of the expression of CD3ε$^+$ (large-sized group), CD8α$^+$ (medium-sized group), IFNγ$^+$ (smallsized group) and CD161 (very-small-sized group) is shown. [Color figure can be viewed at wileyonlinelibrary.com]

comparatively displayed. A selection of them is shown in Figure 2E; namely, CD3ε (general T cell marker), CD8α (combined with CD3, a generic cytotoxic T cell marker), IFN-γ (cytokine produced by subsets of natural killer (NK) cells, NK-T cells (NKT), cytotoxic [CD8$^+$], and helper [CD4$^+$] T cells) and CD161 (cell surface marker expressed by most NK and NKT cells, but also conventional T cell subsets). The first example in Figure 2E shows a large population such as T cells (CD3ε$^+$). As expected from Figure 2D, the density of CD3ε$^+$ cells in the randomized embeddings features seemingly modest fluctuations compared to the non-randomized one. However, in the maximal randomization setting subsections of high protein abundance are difficult to locate by visual inspection. This happens because the injection of noise widens the statistical distributions of abundance values causing the topological arrangement of cells in high dimensions to become more homogenous. Therefore, fewer T cell subsets (regarding CD3 expression levels) can be visually defined from this low dimensional embedding, although the overall structure to generally assign "T cells" by being "positive" for CD3 (irrespective of the expression level) is similar in both randomized and non-randomized cases. The second example illustrates the impact on the delineation of cytotoxic T cells (CD8α$^+$CD3ε$^+$). Here, a subset of cells from this population appears to progressively detach, ultimately becoming a distant neighborhood in the maximal case implying that these cells, in high dimensions, share more characteristics with another cell type. In many cases, such distinction is desirable and the co-expression of IFN-γ and CD161, shown in the next two examples, suggests that NKT cells may be one of these subpopulations. However, the progression of the phenotypic relationships between immune cell types is not maintained here. In addition, the overlap between the respective figures indicates that not all cells belonging to the distant CD8α$^+$ neighborhood are CD8 NKT cells. In fact, a closer inspection shows that this neighborhood is a mixture of CD8 NKT cells and cells from other dimmer subpopulations that are shut away. These complications do not appear in the non-randomized case where a smooth phenotypic transition of CD8 T cells is shown. A similar smooth transition is also shown in Type 1 and Type 2 results. However, in the Type 1 case a subset of IFN-γ expressing cells appears to have been separated from the rest. Another example where the study of dim lymphocyte populations is complicated is shown by the distribution of CD161. In the non-randomized case, as well as Type 1 and Type 2 case, CD161$^+$ cells appear as a phenotypically connected set of subpopulations extending to CD3/CD8 expressing cells and CD3 expressing and IFN-γ producing cells. In the maximal case, however, CD161$^+$ cells appear, again, as two distant neighborhoods whose interpretation is visually problematic.

Randomization Effect on Data Phenotyping

Figure 3 depicts the output of SPADE when the surface marker data are employed and the desired number of clusters is set to 300. The calculated values for the ARI are 0.124, 0.121, and 0.06 for Type 1, Type 2 and maximal randomization while for the average F_1-measure are 0.243, 0.237 and 0.13, respectively. Such low values indicate large discrepancies between the randomized and non-randomized input results. Because these low values may be somewhat expected due to over-clustering, we also checked the effect of the number of clusters on this metrics. Supporting Information Figure S5 illustrates that in the range of 50–500 clusters both indexes monotonically decrease starting with an F_1 score value of about 0.5 and for the ARI, 0.4. These results suggest that randomization has a large effect on SPADE clustering.

As an example, Figure 3 illustrates the effect of randomization on SPADE analysis. On a cell population level, Figure 3A displays the topological distribution of clusters for five color-coded populations that we delineated from the marker expressions shown in Supporting Information Figures S6–S9; that is, CD3ε$^+$CD8α$^-$ and CD3ε$^+$CD8α$^+$ T cells; CD19$^+$ B cells; CD11c$^+$HLA-DR$^+$CD3ε$^-$ Monocytes/Macrophage/Dendritic cells; and CD161$^+$CD3ε$^-$HLA-DR$^-$ NK cells. Each column illustrates the cluster topology for each subpopulation when the non-randomized and randomized data are used, respectively. Even in the case of just these five broader populations, the F_1 measure and ARI range between 0.94 and 0.87 (Type 1 randomization) to 0.87 and 0.77 (maximal randomization), respectively. Beside the discrepancies indicated by the ARI and average F_1-measure, several contradictions are also visible by eye between the tree topologies. Particularly, CD8α$^+$ T cells appear phenotypically uncorrelated to the rest of the T cells (CD3ε$^+$CD8α$^-$) when the maximally randomized data are utilized, but not when the non-randomized or Type 1- and Type 2-randomized data are used. Interestingly, after Type 2 randomization the CD8α$^+$ T cells are split into two CD3ε$^+$ branches. The fact that the right branch is closer to the CD161$^+$CD3ε$^-$HLA-DR$^-$ NK cells suggests that some of these clusters may denote CD8$^+$ NK T cells. Such phenotypically correct case, is not shown in the non-randomized or Type 1 randomized case. On the other hand, the number of clusters attributed to B cells (CD19$^+$) and Monocytes/Macrophage/Dendritic cells (CD11c$^+$HLA-DR$^+$) are becoming drastically less as Type 1, Type 2 and maximal randomization are applied, in favor of T cells. This suggests that the chances of identifying interesting, rare subpopulations between the latter two populations may become lower when randomization is applied on the data. To examine this, we show in Figure 3B the distribution of the expression of a major marker related to each subpopulation. As expected from Figure 3A, the resolution of B cells (CD19$^+$) and Monocytes/Dendritic cells (CD11c$^+$) into subpopulations has become particularly lower after randomization. Regarding CD11c$^+$ cells, in particular, one can potentially distinguish Monocyte-derived Dendritic cells from other Monocyte/Macrophage populations that also express this marker. When, however, maximal randomization is applied this becomes infeasible. Finally, in the case of dim markers like CD161 the effect of randomization has an even stronger impact as the levels of noise injected on its values induces a lot more clusters all over the SPADE tree to have comparable expression values compared to the original case, particularly after the

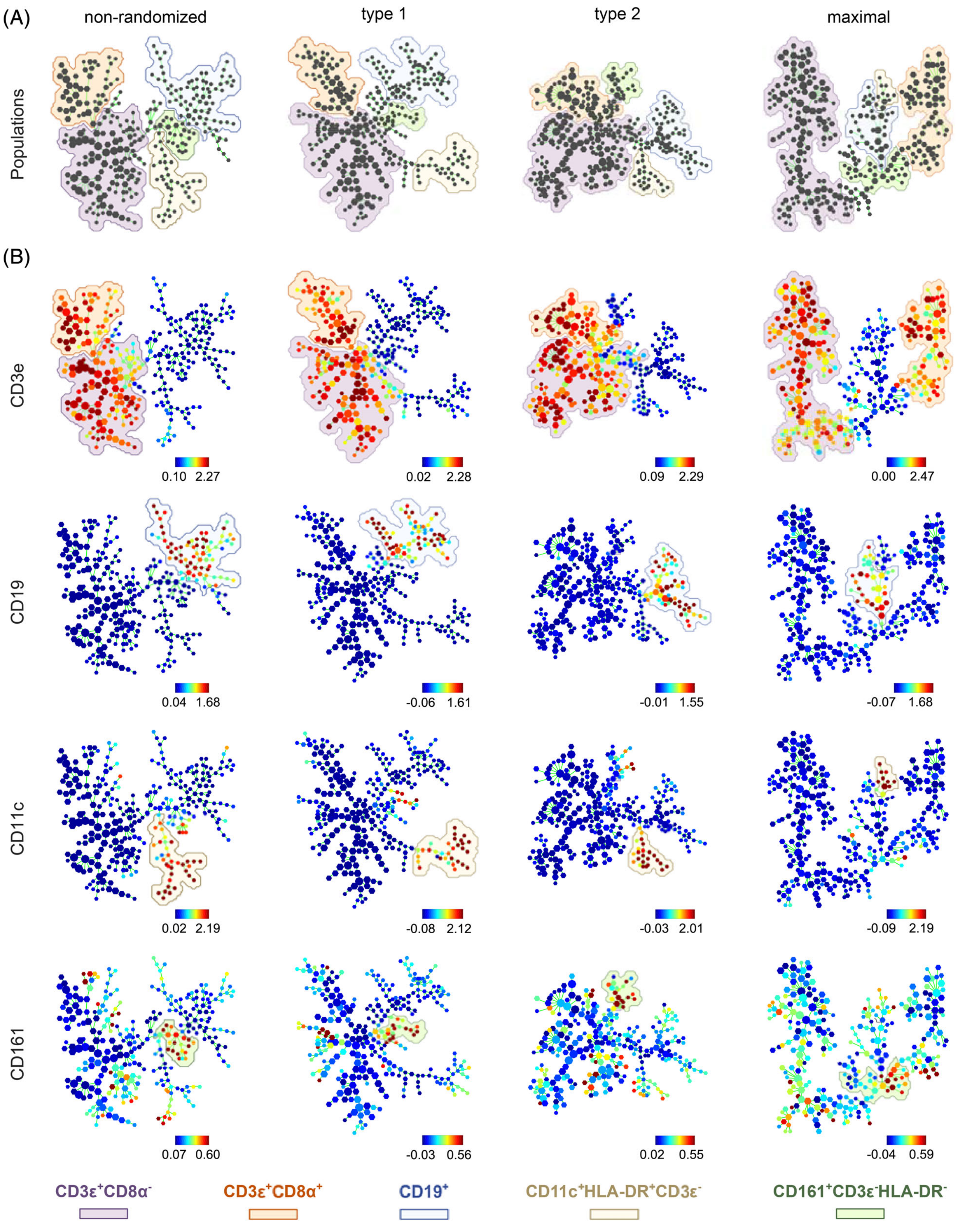

Figure 3. Effect of randomization on SPADE analysis. Each subplot depicts five cell populations: T cells, CD3ε$^+$ CD8α$^-$ (purple) and CD3ε$^+$ CD8α$^+$ (orange); B cells CD19+ (blue); monocytes/macrophage/dendritic cells CD11c$^+$HLA-DR$^+$CD3ε$^-$ (yellow); and NK cells CD161$^+$CD3ε$^-$HLA-DR$^-$ (green). The first row (A) illustrates the cluster topology for each subpopulation when the non-randomized and randomized (Type 1, Type 2, and extreme) data are used, respectively. In (B) each row shows the distribution of the expression of a major marker related to each subpopulation. [Color figure can be viewed at wileyonlinelibrary.com]

maximal randomization. This, shows that the identification of CD161$^+$ cell subsets can become very complicated.

To illustrate this effect more quantitatively on a per marker basis, we matched the clusters between the two results and calculated the $\log_2$ fold-change in average marker expression per matched cluster. Supporting Information Figure S10 illustrates that the average abundance difference may vary between one and eight on the $\log_2$ scale for many clusters and markers in all randomization cases; especially for CD11c, CD8a, CD3e, CD27, and CTLA-4. Similarly, to the previous results, Type 1 randomization seems to have the less significant effect. These results further suggest that randomization has a major effect on the SPADE clustering results.

As a second example of the effect of randomization on cell phenotyping we performed meta-clustering with SOM on all lineage markers, as in SPADE. According to (37) the strategy is, first, to perform SOM and over-cluster cells around 100 grid points and, then, stratify them into 20 meta-clusters. Figure 4B shows the distributions of the F_1-measure and ARI after performing 100 repeats of this procedure. These results show that the correspondence between the non-randomized data and the Type 1, Type 2 and maximally randomized ones progressively drops. Particularly, we found that the average ARI was 0.83, 0.75, and 0.72 and the average F_1-measure was 0.89, 0.80, and 0.78 after applying Type 1, Type 2 and maximal randomization, respectively. Again, this shows that data clustering analysis results may have strong differences.

To assess the meta-clustering further, we overlaid the results on the tSNE plots shown in Figure 2D,E. Each point in Figures 4A is color-coded according to matched meta-clusters. As before, the geometry and topology of cells in large-sized clusters are somewhat similar whereas smaller-sized clusters show several discrepancies. As expected, the non-randomized data generate clusters that can be visually separated from each other conveniently allowing the exploration of more cell subpopulations than in the randomized case. The richness of biological information enclosed in the non-randomized data is particularly shown by the fact that cells in clusters 5 and 10 are each split into several sub-clusters at distant neighborhoods implying that they may enclose cells with somewhat different marker profiles. In contrast, these sub-clusters start converging to a single cluster after randomization. For example, cluster 5 initially shows three subclusters that become two in the Type 1 randomized case and one in the rest. These mean that the original clusters capture information about more than one cell subset and, hence, increasing the number of metaclusters may leverage potential subpopulations. Furthermore, cluster 5 encompasses 7.84% of the total number of cells in the non-randomized case while,

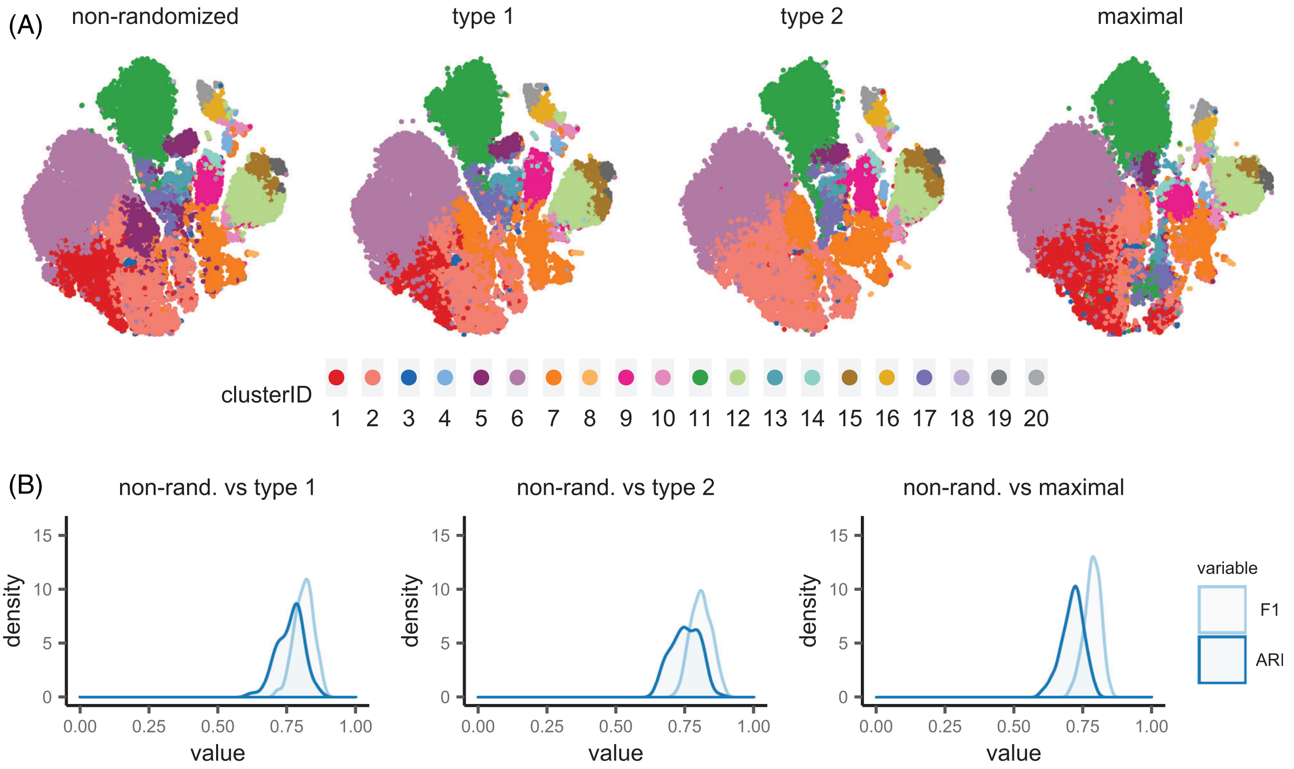

Figure 4. (A) Depicts the clustering result using meta-clustering with SOM when the non-randomized and randomized data sets are employed, respectively. Meta-clusters between the results are matched and each point in the graphs has been color-coded accordingly. Some clusters (e.g. 5) are spread into more than one well-separated neighborhoods showing that this particular cluster contains information about more than one cell subset. Such observation implies that the selected data stratification is not strong enough to capture the full range of cell subsets. (B) Shows the distribution of the F_1-measure and ARI after 100 repeats. [Color figure can be viewed at wileyonlinelibrary.com]

Table 1. Results of the network analyses. Different methods were applied for deriving networks from CyTOF data in two different configurations, namely, non-randomized (NR) and randomized (R). For each case, method and configuration the number of detected edges is shown. Differences between results in the two configurations are summarized as the total number of different edges, as well as the number of edges found using one configuration but not in the other. Differences between networks are also quantified using the structural hamming distance (SHD) metric (see text for details on this metric)

RANDOMIZATION	METHOD	NR EDGES	R EDGES	DIFFERENT EDGES	NR EDGES NOT IN R	R EDGES NOT IN NR	SHD
Type 1	RNs	195	192	3	3	0	3
	GLasso	167	163	4	4	0	4
	rFCI	172	162	38	24	14	88
Type 2	RNs	195	162	33	33	0	33
	GLasso	167	151	22	19	3	22
	rFCI	172	154	46	32	14	109
Maximal	RNs	195	133	62	62	0	62
	GLasso	167	150	31	24	7	31
	rFCI	172	172	64	32	32	132

subsequently, it drops to around 2% after randomization (see Supporting Information Fig. S11). This shows that randomization causes the misplacement of cells into other confounded, large-sized clusters.

To evaluate the robustness of the effects of randomization, we replicated two maximally randomized data sets and performed both SPADE and meta-clustering analyses as before. Selected results are shown in Supporting Information Figure S14. In general, the F_1 score and ARI values were consistent among all replicates. Accordingly, SPADE and meta-clustering outputs were robust to the randomness of the artificial noise. Interestingly, the SPADE result using one

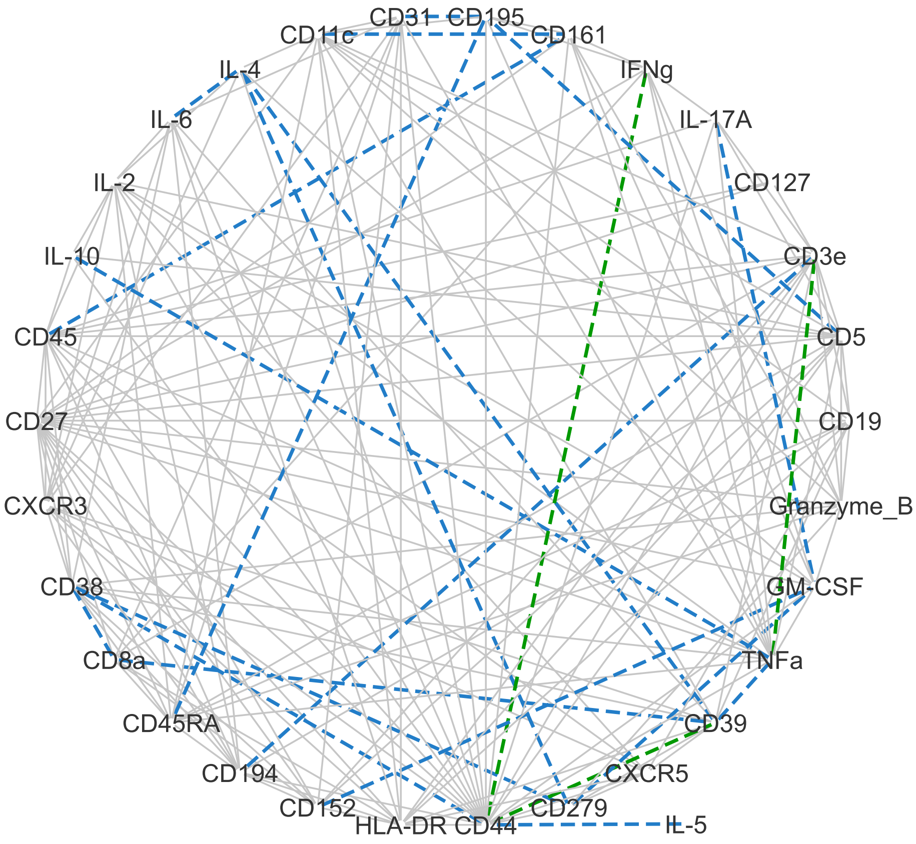

Figure 5. Differences across the networks identified by GLasso using the non-randomized and Type 2 randomized configuration. Edges found in both configurations are reported as solid, gray lines. A blue dashed line indicates an edge found in the non-randomized data but not in the randomized ones; vice versa for the green dashed lines. [Color figure can be viewed at wileyonlinelibrary.com]

replicate, correctly shows only two small subpopulations of CD3ε$^+$CD8α$^-$ and CD3ε$^+$CD8α$^+$ cells to be separated toward NK cells CD161$^+$CD3ε$^-$HLA-DR$^-$, in contrast to the other replicate and the main result shown on Figure 3. This further illustrates the uncertainty induced by randomization on multivariate cell phenotyping results.

Randomization Effect on Network Analysis

We also evaluated the effect of randomization on network structure learning by employing the RNs, GLasso and rFCI algorithms. Table 1 reports the results obtained by applying each network reconstruction method on the non-randomized and randomized CyTOF data configurations. The artificially added noise obfuscates pairwise associations, leading the RN method to identify in the randomized data only a subset of the edges identified in the non-randomized data. The more intense the noise, the higher the number of edges missed in the randomized data. A similar trend is observed also for the other two network analysis methods. Notably, randomization not only obfuscates pairwise correlations but also creates spurious associations in both GLasso and rFCI networks, as indicated by the presence of edges detected in the randomized data but not in the non-randomized ones. Finally, the SHD is considerable higher than the number of different edges for the rFCI networks, indicating a substantial number of edges with wrong orientation in the networks from randomized data.

Figure 5 graphically depicts the differences between the GLasso networks identified in the non-randomized data and in the randomized (Type 2 default) data. Common edges are depicted with solid gray lines, while blue dashed lines indicate edges identified only in the non-randomized data and green dashed lines edges identified only in the randomized data.

Discussion

Randomization Is Largely Overlooked by both Data Analysts and Method Developers

CyTOF is a disruptive technology with great potentials but as in any novel cell measuring method the standardization of data quality control is a slow process. It is, however, extremely important to reach to a consensus soon because only in this way good quality data will be generated and, subsequently, good quality computational algorithms will be developed. Unfortunately, such discussion has attracted only little attention. In fact, most CyTOF data prepreprocessing tasks are currently heuristic with only few of them being based on rigorous mathematical grounds. For example, cell data generation from raw measurements is based on arbitrarily selected switchover thresholds (shown in Fig. 1B). Similarly, data transformation is based on an arbitrary chosen co-factor that defines the range of values over which the data are linearly or logarithmically rescaled, neglecting how this may affect the statistical properties of each of the underlying marker distributions. On the other hand, data normalization has received some systematic investigation; however, the choices for the analyst are limited to only two methods that assume the noise is linear. Clearly, CyTOF data preprocessing is still in its infancy and focused research on these issues is urgently required.

Our main focus has been on randomization because it is the only unjustifiable task for computational analysis. What is alarming about randomization is that it is a heuristic borrowed “as is” from flow cytometry data analysis practices and is being applied by default by the device software. Despite the fact that the software allows reprocessing of the raw data files to remove or adjust various kinds of preprocessing, the proper use of randomization has received attention by only a small part of the community. Particularly, the effects of randomization to CyTOF manual gating were discussed in a Cytoforum (http://cytoforum.stanford.edu) post back in 2013 and also, briefly, in 2014 (38,39). Then, another post from 2015 in the same forum discussed about the problems caused by the application of data randomization to multidimensional analysis (40). However, the discussion did not reach to a clear consensus rather to some arbitrary recommendations without any strong proof. The general lack of awareness that probably exists until today is seen in a more recent post from 2017 where the issue is brought up again because a clustering analysis generated some uninterpretable cell populations (41). What is even more alarming is that despite these discussions over the years the vast majority of high-dimensional algorithms in the field have been developed based on randomized data. This is reflected by our literature survey for the most cited works that proves that public data sets have been randomized in one way or another. One could argue that the randomized abundances could be converted back to their original discrete values if the randomization scheme is known. In this way, the algorithms’ robustness against randomization can be evaluated. To the best of our knowledge, however, besides the normalization method and the rescaling transform employed, the exact randomization settings with which the publicly available data sets were preprocessed or the assumptions under which the use of randomized data are justified, are nowhere reported.

Randomization May Significantly Affect Downstream High-Dimensional Analyses

We have demonstrated the effect of randomization on a wide range of algorithms that are typically used for data visualization, cell phenotyping, and the reconstruction of protein networks. In principle, randomization should have no noticeable effect on bivariate visualizations used typically during gating, other than alleviating the striation patterns. Accordingly, no appreciable effect on descriptive statistics is expected assuming that the number of data samples is large (e.g., number of cells >10 K). For instance, assuming a large number of cells Type 1 randomization will shift the data by an average of −0.5 while Type 2 randomization will have no significant effect. In contrast, as the number of cells drops, significant discrepancies may appear. Our findings after applying three different randomization strategies pointed out several significant discrepancies in every analysis method when the non-randomized or randomized data are used as input. Since any two configurations differ only on whether the randomization was performed or not, any difference found in the results has been due solely to the noise artificially introduced in the data and does not reflect any underlying biological mechanism. As expected, the Type 1 randomization we employed conferred the smallest effect among the three cases we examined while the maximal randomization case the largest ones. Regarding maximal randomization, we should emphasize that Supporting Information Figures S1 and S2 clearly depict that there is no consensus neither as to what type of randomization is typically used nor as to how many times it is applied during data preprocessing. Therefore, even if the maximal randomization is not default as Type 1 is, it stands as a reasonable case to illustrate the limits of the potential effects.

Our results from linear dimensionality reduction using PCA showed that the effect of randomization in data visualization could be somewhat small. However, a more detailed analysis with JIVE revealed that the useful biological variance of the data becomes considerably corrupted after any randomization. In contrast to PCA, an MDS plot between the median abundance values illustrated differences between not only the non-randomized and the randomized cases, but also between the different randomization strategies. Accordingly, tSNE nonlinear analysis indicated considerable effects of randomization on the topology of the low dimensional embeddings. On top of that, it showed that dim cell markers and small-sized cell groups could be largely affected by this heuristic rendering the stratification of the data into several cell subsets difficult and the subsequent detection of rare cell subpopulations potentially improbable. This fact is further supported by the output of cell phenotyping using clustering. SPADE analysis, in particular, showed that extensive randomization might severely confound the interpretation of the output by corrupting the resolution of the data. Of note, is the contradiction between the different randomization strategies where the CD8α^+ T cells get split into two branches raising important implications regarding the recovered phenotypical progression of cells. Similarly, meta-clustering revealed that randomization may conceal important phenotypic information required for further stratification of cells. Furthermore, because the output of multivariate analyses can be sensitive to random starts, we employed the deterministic version of SPADE and specified the same random seed during meta-clustering. However, randomization is by itself stochastic and, hence, the output can also be sensitive to the randomness of the added noise. Our results on replicate data sets have shown consistent randomization effects both on the level of F_1 score and ARI and the potential misinterpretation of multivariate cell phenotyping analysis.

Along the same lines, the networks identified from randomized and original data are quite different. This is expected because the negative effect of measurement error on network reconstruction is well-known in the literature (42). The most apparent effect is the attenuation of associations, resulting in a lower number of detected edges in the randomized data across all methods (Table 1). Another, more subtle effect is the possible inflation of partial correlations. Two measurements A and B may show a significant univariate association $|\rho_{A,\ B}|>0$ which disappears when conditioning on a set **C** of other variables, that is, $\rho_{A,\ B\mid C}=0$. If measurement error is added to the variables, the set **C** may not convey enough information for explaining the correlation between A and B, and a spurious correlation $|\rho_{A,\ B\mid C}|>0$ may be detected (43). In our experiments this effect is a probable explanation for the edges detected in the randomized data but not in the non-randomized by the GLasso and rFCI methods. Some specific methods for correcting the effect of measurement errors in network reconstruction have been recently proposed (44,45) however avoiding the artificial introduction of this type of noise remains of paramount importance.

Recommendations

We here have described current CyTOF data preprocessing methodologies with emphasis on randomization. We showed that this task corrupts the data in ways such that they lead to unreliable biological inferences. The biological results we have shown do not necessarily apply to all CyTOF data sets. Rather, they are indicative of the minor and major implications that randomization may confer on the multidimensional data analyses. On such basis, our recommendation is to avoid as much as possible the use of randomization when analyzing CyTOF data or when developing new multivariate analysis tools. In general, to avoid potential pitfalls the systematic study of all preprocessing steps is recommended. Another recommendation we leave as future work, is to evaluate the effect of randomization relative to sample-to-sample and lab-to-lab variability (46). This variability is independent of the noise injected by randomization. However, it would be useful to examine the combined effect on the data and provide means to control for it. Furthermore, a detailed assessment of the steps prior to cell data (.FCS) processing is also required in order to evaluate how standard practices may be affecting downstream analysis.

Regarding the development of tools, we also recommend that these should be taking into account the fact that CyTOF data are initially count data or at least discrete and, so, processing should take place under certain mathematical assumptions. Finally, the importance of open-data policies in life science is widely acknowledged, and their positive impact is provably relevant (47). Thus, we recommend that scientific works using CyTOF data should always provide their raw data, along with a clear description of all preprocessing steps used for their analysis, in order to ensure replicability, re-usability and the correctness of future analysis. Scripts and data files to reproduce the preceding results are available from GitHub (https://github.com/mensxmachina/CyTOF-Randomization) and FlowRepository (repositories FR-FCM-Z2ZW, FR-FCM-Z24G, FR-FCM-Z28H).

Acknowledgment

The authors acknowledge the SciLifeLab National Mass Cytometry Facility services in Stockholm for performing mass cytometry; particularly its members doctors. Tadepally Lakshmikanth, Yang Chen, Jaromir Mikes, and Petter Brodin for the many helpful discussions and feedback regarding sample preparation protocols and CyTOF data generation settings. In addition, the authors would like to thank the anonymous referees for their insightful comments and key suggestions during the review of this manuscript. The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ ERC Grant Agreement no. 617393; CAUSALPATH – Next Generation Causal Analysis project. Funding for open access charge: ERC.

Literature Cited

1. Bandura DR, Baranov VI, Ornatsky OI, Antonov A, Kinach R, Lou X, Pavlov S, Vorobiev S, Dick JE, Mass Cytometry TSD. Technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. Anal Chem 2009;81:6813–6822.
2. Spitzer MH, Mass Cytometry NGP, Cells S, Features M. Mass cytometry: Single cells, many features. Cell 2016;165:780–791.
3. Saeys Y, Van GS, Lambrecht BN. Computational flow cytometry: Helping to make sense of high-dimensional immunology data. Nat Rev Immunol 2016;16:449–462.
4. Diggins KE, Ferrell PB, Irish JM. Methods for discovery and characterization of cell subsets in high dimensional mass cytometry data. Methods 2015;82:55–63.
5. Weber LM, Robinson MD. Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. Cytom Part A 2016;89:1084–1096.
6. Bruggner RV, Bodenmiller B, Dill DL, Tibshirani RJ, Nolan GP. Automated identification of stratifying signatures in cellular subpopulations. Proc Natl Acad Sci U S A 2014;111:E2770–E2777.
7. Arvaniti E, Claassen M. Sensitive detection of rare disease-associated cell subsets via representation learning. Nat Commun 2017;8:1–10. Available at. https://doi.org/10.1038/ncomms14825.
8. Newell EW, Cheng Y. Mass cytometry: Blessed with the curse of dimensionality. Nat Immunol 2016;17:890–895.
9. Li YH, Li D, Samusik N, Wang X, Guan L, Nolan GP, Wong WH. Scalable multisample single-cell data analysis by partition-assisted clustering and multiple alignments of networks. PLoS Comput Biol 2017;13:e1005875.
10. Triantafillou S, Lagani V, Heinze-Deml C, Schmidt A, Tegner J, Tsamardinos I. Predicting causal relationships from biological data: Applying automated causal discovery on Mass Cytometry data of human immune Cells. Sci Rep 2017;7:12724.
11. Cosma A. A time to amaze, a time to settle down, and a time to discover. Cytom Part A 2015;87:795–796.
12. Leipold MD. Another step on the path to mass cytometry standardization. Cytom Part A 2015;87:380–382.
13. Finck R, Simonds EF, Jager A, Krishnaswamy S, Sachs K, Fantl W, Pe'er D, Nolan GP, Bendall SC. Normalization of mass cytometry data with bead standards. Cytom Part A 2013;83A:483–494.
14. Tricot S, Meyrand M, Sammicheli C, Elhmouzi-Younes J, Corneau A, Bertholet S, Malissen M, Le Grand R, Nuti S, Luche H, et al. Evaluating the efficiency of isotope transmission for improved panel design and a comparison of the detection sensitivities of mass cytometer instruments. Cytom Part A 2015;87:357–368.
15. Ornatsky OI, Kinach R, Bandura DR, Lou X, Tanner SD, Baranov VI, Nitz M, Winnik MA. Development of analytical methods for multiplex bio-assay with inductively coupled plasma mass spectrometry. J Anal At Spectrom 2008;23:463–469.
16. Roederer M. Spectral compensation for flow cytometry: Visualization artifacts, limitations, and caveats. Cytometry 2001;45:194–205.
17. Papoutsoglou G, Athineou G, Lagani V, Xanthopoulos I, Schmidt A, Éliás S, Tegnér J, Tsamardinos I, SCENERY: A web application for (causal) network reconstruction from cytometry data. Nucleic Acids Res 2017;45:W270–W275.
18. Ornatsky O, Bandura D, Baranov V, Nitz M, Winnik MA, Tanner S. Highly multiparametric analysis by mass cytometry. J Immunol Methods 2010;361:1–20.
19. Zunder ER, Finck R, Behbehani GK, Amir E-AD, Krishnaswamy S, Gonzalez VD, Lorang CG, Bjornson Z, Spitzer MH, Bodenmiller B, et al. Palladium-based mass tag cell barcoding with a doublet-filtering scheme and single-cell deconvolution algorithm. Nat Protoc 2015;10:316–333.
20. Finak G, Perez JM, Weng A, Gottardo R. Optimizing transformations for automated, high throughput analysis of flow cytometry data. BMC Bioinformatics 2010; 11:546.
21. Shaham U, Stanton KP, Zhao J, Li H, Raddassi K, Montgomery R, Kluger Y. Removal of batch effects using distribution-matching residual networks. Bioinformatics 2017;33(16):2539–2546.
22. Chevrier S, Crowell HL, Zanotelli VRT, Engler S, Robinson MD, Bodenmiller B. Compensation of signal spillover in suspension and imaging Mass Cytometry. Cell Syst 2018;6:612–620.e5.
23. Van Der Maaten LJP, Hinton GE. Visualizing high-dimensional data using t-sne. J Mach Learn Res 2008;9:2579–2605.
24. Lock EF, Hoadley KA, Marron JS, Nobel AB. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. Ann Appl Stat 2013; 7:523–542.
25. Qiu P. Toward deterministic and semiautomated SPADE analysis. Cytom Part A 2017;91:281–289.
26. Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. Proc Natl Acad Sci U S A 2000;97:12182–12186.
27. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics 2008;9:432–441.
28. Colombo D, Maathuis MH, Kalisch M, Richardson TS. Learning high-dimensional directed acyclic graphs with latent and selection variables. Ann Stat 2012;40: 294–321.
29. Zhang J. Causal reasoning with ancestral graphs. J Mach Learn Res 2008;9: 1437–1474.
30. Malinsky D, Danks D. Causal discovery algorithms: A practical guide. Philos Compass 2018;13:e12470.
31. Hubert L, Arabie P. Comparing partitions. J Classif 1985;2:193–218.
32. Samusik N, Good Z, Spitzer MH, Davis KL, Nolan GP. Automated mapping of phenotype space with single-cell data. Nat Methods 2016;13:493–496.
33. Tsamardinos I, Brown LE, Aliferis CF. The max-min hill-climbing Bayesian network structure learning algorithm. Mach Learn 2006;65:31–78.
34. Triantafillou S, Tsamardinos I. Score-based vs Constraint-based Causal Learning in the Presence of Confounders. In: Eberhardt F, Bareinboim E, Maathuis MH, Mooij JM, Silva R, editors. Proc. UAI 2016 Workshop on Causation: Foundation to Application co-located with 32nd Conf. Uncertain. Artif. Intell. (UAI 2016). Vol 1792. {CEUR} Workshop Proceedings. Jersey City: CEUR-WS.org; 2016. pp 59–67. Available at: http://ceur-ws.org/Vol-1792/paper7.pdf.
35. Olin A, Henckel E, Chen Y, Lakshmikanth T, Pou C, Mikes J, Gustafsson A, Bernhardsson AK, Zhang C, Bohlin K, et al. Stereotypic immune system development in newborn children. Cell 2018;174:1277–1292. -e14.
36. Lakshmikanth T, Olin A, Chen Y, Mikes J, Fredlund E, Remberger M, Omazic B, Brodin P. Mass Cytometry and topological data analysis reveal immune parameters associated with complications after allogeneic stem cell transplantation. Cell Rep 2017;20:2238–2250.
37. Nowicka M, Krieg C, Weber LM, Hartmann FJ, Guglietta S, Becher B, Levesque MP, Robinson MD. CyTOF workflow: Differential discovery in high-throughput high-dimensional cytometry datasets. F1000Research 2017;6:748.
38. Smith R. CyTOF .fcs files and proper gating. http://cytoforum.stanford.org 2013. Available at: http://cytoforum.stanford.edu/viewtopic.php?f=3&t=97&p=246&hilit=noise#p246.
39. Dawson N. Odd results? Trouble analyzing. http://cytoforum.stanford.org 2014. Available at: http://cytoforum.stanford.edu/viewtopic.php?f=3&t=124.
40. Stevens C. Negative values in CyTOF. http://cytoforum.stanford.org 2015. Available at: http://cytoforum.stanford.edu/viewtopic.php?f=1&t=268&hilit=randomization&start=10.
41. Khan N. VISNE and zero values in CyTOF data. http://cytoforum.stanford.org 2017. Available at: http://cytoforum.stanford.edu/viewtopic.php?f=3&t=677.
42. Nagarajan R, Scutari M. Impact of noise on molecular network inference. PLoS ONE 2013;8:e80735. Available at. http://dx.plos.org/10.1371/journal.pone.0080735.
43. Lagani V, Triantafillou S, Ball G, Tegnér J, Tsamardinos I. Probabilistic computational causal discovery for systems biology. Part of the Studies in Mechanobiology, Tissue Engineering and Biomaterials book series. Uncertain Biol 2015. 17 p 33–73. Available at: http://dx.doi.org/10.1007/978-3-319-21296-8_3.
44. Hemani G, Tilling K, Davey Smith G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. PLoS Genet 2017;13: e1007081. Available at. https://dx.plos.org/10.1371/journal.pgen.1007081.
45. Blom T, Klimovskaia A, Magliacane S, Mooij JM. An upper bound for random measurement error in causal discovery. 2018. Available at: http://arxiv.org/abs/1810.07973.
46. Leipold MD, Obermoser G, Fenwick C, Kleinstuber K, Rashidi N, McNevin JP, Nau AN, Wagar LE, Rozot V, Davis MM, et al. Comparison of CyTOF assays across sites: Results of a six-center pilot study. J Immunol Methods 2017;453:37–43. Available at. https://doi.org/10.1016/j.jim.2017.11.008.
47. Perez-Riverol Y, Zorin A, Dass G, Glonț M, Vizcaino JA, Jarnuczak A, Petryszak R, Ping P, Hermjakob H. Quantifying the impact of public omics data. Nat. Commun. 2019;10:3512. https://doi.org/10.1101/282517. Available at: http://www.nature.com/articles/s41467-019-11461-w.