Title

cellHarmony: Cell-level matching and comparison of single-cell transcriptomes

Authors

Erica AK DePasquale[1,5], Kyle Ferchen[3], Stuart Hay[1], H. Leighton Grimes[2,3,4], Nathan Salomonis[1,2,5]

[1]Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH

[2]Department of Pediatrics, University of Cincinnati School of Medicine, Cincinnati, Ohio, USA.

[3]Division of Immunobiology and Center for Systems Immunology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA.

[4]Division of Experimental Hematology and Cancer Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH

[5]Department of Biomedical Informatics, University of Cincinnati, Cincinnati, OH

Corresponding Author

Nathan Salomonis

Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

650-576-1646

Nathan.Salomonis@cchmc.org

ABSTRACT

To understand the molecular etiology of human disease, precision analyses of individual cell populations and their molecular alternations are desperately needed. Single-cell genomics represents an ideal platform to enable to the quantification of specific cell types, the discovery of transcriptional cell states and underlying molecular differences that can be compared across specimens. We present a new computational approach called cellHarmony, to consistently classify individual cells from a query (i.e., mutant) against a reference (i.e., wild-type) dataset to discover crucial differences in discrete or transitional cell-populations. CellHarmony performs a supervised classification of new scRNA-Seq data against *a priori* delineated cell populations and associated genes to visualize the combined datasets and derive consistent annotations in a platform-independent manner. Such analyses enable the comparison of results from distinct single-cell platforms against well-curated references or against orthogonal profiles from a related experiment. In addition, cellHarmony produces differential expression results from non-confounded aligned cell populations to explore the impact of chemical, genetic, environmental and temporal perturbations. This approach works seamlessly with the unsupervised classification and annotation of cell-states using the software ICGS in AltAnalyze. Using cellHarmony, we demonstrate novel molecular and population insights in scRNA-Seq data from models of Acute Myeloid Leukemia, across technological platforms and using references derived from the Human Cell Atlas project.

INTRODUCTION

Single-cell RNA-Seq provides the unique ability to profile transcripts from diverse cell populations along a continuum of related or disparate states (Olsson et al., 2016). From a developmental perspective, scRNA-Seq can define molecular decisions that describe the differentiation of progenitors to one or more distinct states. The ability to accurately map such cellular decisions becomes important when considering diseases in which abnormal cell states arise, such as cancer. Although approaches to identify cell populations from scRNA-Seq are abundant, new methods are needed to map individual cells that may underlie discrete or transitional cell populations. The ability to align cells from different distinct conditions, animals or individuals opens the door for automated approaches to compare cells from distinct cell populations to identify population-specific regulated genes and pathways.

Recently, several new approaches have been developed to align genetically similar cells within the same platform, across distinct platforms, and even across species for comparison. These methods include scmap, Seurat CCA, BISCUIT and MNN, which are designed to align similar cells from distinct datasets and account for batch or other technical effects (Butler and Satija, 2017; Haghverdi et al., 2018; Kiselev et al., 2018; Prabhakaran et al., 2016). In the case of Seurat CCA, these analyses provide the capability to derive unsupervised scRNA-Seq subtypes from the combined data to align cells. The software scmap projects queried scRNA-Seq cells against a reference set of classified cell states and cells to obtain cell-type predictions. Recently, we demonstrated the ability to classify cells across scRNA-Seq platforms (Fluidigm, DropSeq, 10x Genomics) to standardize the detection of common cell states in mouse embryonic kidney and compare genes uniquely expressed in those associated cell states (Magella et al., 2017). This approach employed k-nearest neighbor classification of individual cells against reference cell-state centroids and provides an important proof of concept for alignment and comparison of individual cell profiles. While an important step, such approaches are likely limited in their ability to distinguish rare transitioning populations and differences that result from disease.

Herein, we describe a new approach called cellHarmony which provides the unique ability to align, visualize and compare scRNA-Seq data from a query sample (perturbed or non-perturbed) against all referenced cell states. Unlike other approaches to perform *de novo* alignment between different scRNA-Seq datasets or captures, cellHarmony is designed to align cells from a query dataset of single-cell profiles directly to a reference (cells or centroids). This approach can be run in conjunction with results produced from unsupervised or supervised single-cell analyses in the software AltAnalyze, on the command-line or through an intuitive graphical user interface. Beginning with unsupervised clustering results from ICGS, classified query cells

can be arranged along a continuum of discrete, transitional or multi-lineage cell states. In doing so, one can: 1) determine the frequency and specificity of individual cell captures against larger ideally curated reference datasets based on each individual cell's closest match using Pearson correlation and 2) quantify differences in the expression of genes from aligned cell states in two compared single-cell datasets. While the latter should only be performed for comparable datasets analyzed with the same single-cell platform, the alignment-only step can be applied to datasets obtained from different laboratories or experimental platforms to obtain predicted cell states in a supervised manner. These results highlight unique molecular cell-biology in perturbed cells corresponding to distinct cell populations.

METHODS

Algorithm Design: cellHarmony uses a KNN classification approach (K=1) to assign class labels for individual cells in the queried dataset against an established reference set of cell or centroid gene expression profiles. The nearest neighbor of a query cell profile is assigned to its best match in the reference set based on all possible cell-cell Pearson correlations. cellHarmony performs its supervised analysis only on genes which are dynamically expressed from initial unsupervised analyses and thus represent core cell-identity gene expression programs, excluding variable or stochastic programs.

We originally developed this approach to classify cells from different scRNA-Seq technologies in embryonic kidneys according to a common consensus set or reference centroids (Magella et al., 2017). In the current implementation, the class label (e.g., cluster 1) of the nearest neighbor is reported as the label for the queried cell. A Pearson correlation cutoff threshold (0-1, user-defined) is required to exclude cells which insufficiently match to the reference. Such outlier cells may represent cell-types not found in the reference, poorly sequenced cells or multiplet cell profiles (e.g., doublets). Such cells themselves may be of interest, as they may represent novel cell populations missing from the reference. As such, we reported these cells to the user to allow for independent analyses (CellClassification results folder). By default, cellHarmony can use different outputs from the AltAnalyze analysis tool Iterative Clustering and Guide-gene Selection (ICGS), which are formatted in a standard tab-delimited text file format but can also work with results from external workflows (e.g., Seurat) (Butler and Satija, 2017; Olsson et al., 2016). For such analyses, cell-populations that have already inferred identities (e.g., monocyte) are recommended, as those annotations will be later projected onto mapped cells from the query

dataset. Importantly, ICGS results already include predictions for cell-type identity using prior cell-type specific references.

Implementation: cellHarmony is compatible with Python 2.7 and is distributed as a component of the software AltAnalyze (https://github.com/nsalomonis/altanalyze) and thus works seamlessly with other modules in AltAnalyze (e.g., ICGS). It can be run both on the command-line and through the AltAnalyze version 2.1.1 graphical user-interface (Additional Analyses Menu > Cell Classification Menu). Pre-compiled graphical user-interface distributions are provided from http://altanalyze.org and the command-line version from Github or via installation from PyPI (`$ sudo pip install AltAnalyze`). A development R version of cellHarmony performs the basic cell-alignment analyses but does not currently support downstream differential expression or multi-reference merge analyses (https://github.com/EDePasquale/cellHarmony).

Combined Reference Creation: cellHarmony is capable of merging multiple references which may be confounded by batch or donor effects (cellHarmony merge function). While CellHarmony does not explicitly correct for batch effects, when providing multiple ICGS or other unsupervised subtype detection results (i.e., multiple individual donor ICGS results) it will produce a merged result file with the union of all supplied marker genes and averaged similar-cell profiles. The standard inputs for this analysis are the ICGS clustered heatmap text file results or ICGS MarkerGenes heatmap text file (see https://altanalyze.readthedocs.io/). To produce these results, the cellHarmonyMerge function: 1) selects all unique marker genes from the collectively provided set of inputs, 2) imports the expression data for these genes from the original complete gene matrix (e.g., AltAnalyze ExpressionInput “exp.” file) and converts these to log2 values, 3) computes median expression for cells (medoid) within the same identified cluster for all marker genes, 4) averages similar medoid “clusters” based on all pairwise medoid comparisons (Pearson correlation>0.9) to create reference centroids, 5) filters the combined dataset to only include genes with non-zero values for all columns, 6) re-clusters the new reference centroids (HOPACH clustering) to produce a re-ordered matrix of reference cell-population centroids and genes. The resultant reference text file can be used for downstream cellHarmony analyses. As an example, we have applied this workflow to 35 distinct cell populations we identified from bone marrow RNA-

Seq generated by the Human Cell Atlas project, following independent ICGS analyses of each independent donor (Hay et al., 2018).

cellHarmony Outputs: cellHarmony produces as output multiple tabular and visualization results, depending on the type reference dataset supplied (full expression matrix versus pre-filtered). The initial outputs of cellHarmony are: 1) final association z-score matrix derived from the Pearson correlation coefficients for all cells, 2) an expression matrix in which each cell is placed adjacent to its best match, 3) query-only cell matrix with cells ordered and annotated according to the classification, 4) gene expression heatmaps of the expression matrices, 5) cell frequency and gene expression difference bar charts, 6) statistical differences in the frequency of aligned cell populations between reference and query samples, 7) UMAP projection of the query and reference cells combined, and 8) a MarkerFinder (Olsson et al., 2016) ordered heatmap with enriched Gene Ontology terms of fold differences in all compared cell populations (**Fig. 1A**). The log-normalized expression profiles for both the reference and query are displayed as a combined heatmap, in the reference gene and cell order, with the query inserted alongside each cellHarmony match to assess their relative similarity. The frequency of cells present or absent from the query in each cell population are further reported and statistically quantified using a chi-squared test to allow for the assessment of the lineage impact with cellular, molecular or genetic perturbation in the query (**Fig. 1B**).

Differential Expression Analysis: Cell populations with similar or dissimilar frequencies can result in differential expression when compared directly (**Fig. 1C**). When the reference input file (e.g., ICGS heatmap text file) is present in a standard AltAnalyze output directory (e.g., ICGS folder), the software will automatically search for the associated full expression dataset file in the ExpressionInput directory (exp. prefix file - https://altanalyze.readthedocs.io/). Using the cluster labels in the reference file, cells assigned to a specific label are compared in the query to the same labelled group in the reference (e.g., monocytes). All pairwise query and reference comparisons are performed for a given cluster in which at least 4 cells are present in both the query and reference. Differential expression is performed using an empirical Bayes moderated t-test (limma) with Benjamini-Hochberg adjustment. The default threshold for differential expression is fold>1.5 and p<0.05 (FDR corrected). The results are saved to tab-delimited text files with summary statistics and basic annotations (*DifferentialExpression_Fold_1.5_adjp_0.05* directory) along with a summary graphical output (see **Fig. 4F**). These differentially expressed genes are subsequently compared across all comparisons to identify genes with higher specificity for

specific comparisons using MarkerFinder and output to a final summary heatmap with statistically enriched Gene Ontology terms (GO-Elite algorithm) for MarkerFinder assigned genes to each comparison (see **Fig. 4G** and the example data provided with the software).

If using external outputs from tools such as Seurat, differential expression can be computed between the cellHarmony assigned cell populations in those tools (e.g., MAST differential expression) or within AltAnalyze (metaDataAnalysis function), after merging the query and reference expression files (see cellHarmony documentation).

Algorithm Evaluation: The psych R package was used to calculate inter-rater agreement using Cohen's kappa (unweighted) to evaluate the performance of cellHarmony in terms of ability to place originally sampled cells back into their source populations. CellHarmony was tested using variable genes selected by ICGS or MarkerFinder. scmap was tested using ICGS genes and scmap variable genes to evaluate the impact of feature selection. In brief all 23,956 genes were input to scmap with 30% of the cells extracted from each ICGS cluster (with a minimum of 1 for each group) for the testing set. Of the remaining 70% of the cells, features were selected using the default expression and dropout method available in scmap to select 500 genes as variable features. The model was projected onto the test set.

An option in cellHarmony is the ability to specify a minimum correlation required to classify a cell as belonging to one of the present cell clusters. To evaluate Type I errors, one distinct cell cluster was excluded from each training dataset and added to the testing set (10% sampled from the remaining clusters), with cells falling below the correlation cutoff saved in a separate outlier cell table instead of being aligned to the reference. Correlation cutoffs between 0 and 1 in 0.05 increments were tested to determine which correlation reduced errors in placement into either the training set or outlier cell table. This second testing was performed on each individual cluster in the testing datasets to assess consistency of the method across varying similarities of query cells and Cohen's kappa was used as the error metric. The evaluation script for cellHarmony for different datasets can be found in the cellHarmony R Github repository. Putative cell multiplet profiles were removed in the human Melanoma scRNA-Seq dataset using a prototype version of DoubletDecon (version 1.0 alpha).

Differential gene expression estimates from the software cellHarmony (empirical Bayes), SCDE and MAST were compared using bulk RNA-Seq as a control. Bulk RNA-Seq T-cells and B-cells (GSE51984) were aligned to the human genome (hg19) with the program STAR and analyzed using AltAnalyze to identify differentially expressed genes (DEGs) with an FDR corrected $p<0.05$. Single-cell RNA-Seq from human PBMCs was downloaded from the 10x

Genomics website (https://support.10xgenomics.com/single-cell-vdj/datasets) and processed in AltAnalyze with the ICGS algorithm to identify a CD8+ T-cell population and B-cell population. These data were compared in SCDE, MAST and AltAnalyze. For genes identified in the bulk and single-cell comparisons with a fold change in the same direction, DEGs were compared to calculate sensitivity and specificity.

RESULTS

Cell-Alignment Performance

cellHarmony is designed to be compatible with cell-to-cell or cell-to-centroid assignments. To assess the global accuracy of the cellHarmony in matching cells to their most similar cell, Monte Carlo cross-validation was performed. The core matching module of cellHarmony was recreated in R to facilitate the testing process, which is described as follows. As an evaluation dataset we used a previously described Fluidigm scRNA-Seq dataset of 382 murine bone marrow hematopoietic progenitors (BM), which includes reasonably well-defined rare and common progenitors as well as transitional cell populations. We additionally evaluated an index-sorted human Melanoma dataset which combines over 4,000 cells from 19 different patients, with distinct immune and non-immune cell populations (Melanoma).

For each test dataset, 10% of the cells from each empirically derived cell cluster (ICGS determined) were randomly selected as the testing set and removed from the remaining cells, which became the training set. cellHarmony was used to place test cells back into the training expression matrix (adjacent to the classified cell). As similarly evaluated in the scmap publication, we used a combination of unweighted Cohen's kappa, to assess inter-rater agreement while accounting for chance agreement, and percent assignment. The average of these scores comprised the performance metric, with high performance values indicating correct placement in the original groups. This first test was performed over 10 rounds of simulation for each validation dataset. For the Melanoma dataset we additionally removed prospective hybrid cell profiles (aka multiplets) using a recently described doublet detection method (DoubletDecon) (DePasquale et al., 2018).

The Cohen’s kappa analysis indicates that cells can be placed with a high degree of inter-rater agreement (kappa of 0.76 in the BM and 0.94 in Melanoma), even without the use of an explicit correlation cutoff (100% assignment of tested cells) (**Fig. 2A-D**). Increasing the correlation cutoff beyond a Pearson rho=0.7, increased the Cohen’s kappa to 0.85 (BM) and 0.97 (Melanoma), respectively, while retaining 62% and 90% of cells, respectively. Without excluding

putative doublets in Melanoma, kappa decreased to 0.84 with no correlation cutoff and to 0.92 with a correlation cutoff of 0.7 (**Fig. S1A**).

When we examine cells that are inaccurately aligned to clusters, mis-aligned cells most often were assigned to adjacent highly similar clusters (**Fig. 2E, Fig. S1B**). In the BM dataset, two highly related hematopoietic stem cell (HSC) progenitor clusters (HSCP-1, HSCP-2) were found to account for 43% misplacements. Likewise, two clusters (c7 and c8) resulted in the greatest number of misplacements in the Melanoma scRNA-Seq dataset. Here, clusters 2, 7 and 8 correspond to T-cells. While cluster 2 is *CD8A*+ and clusters 7 and 8 are *CD4*+ by RNA, cluster 7 appears to represent lower quality *CD4*+ T-cell transcriptomes, with no distinguishing marker genes by MarkerFinder. Hence, it appears that misplacements are not always inaccurate but rather can reflect developmental intermediates or imprecise assignment of cells by ICGS.

Given that cells from a query dataset may not be represented in the reference, due to selection strategy or disease-state, we additionally excluded one cell cluster from each evaluation dataset and tested the placement of these cells in the reference (type I errors). Iterative removal of all cells in each cell-cluster, alongside removal of 10% of the cells from the remaining clusters, and testing their classification resulted in variable performance depending on the similarity of the removed cluster to the remaining reference clusters (**Fig. 2F**, **Fig. S1C**). For example, below a correlation cutoff of 0.7, removal of monocytic or granulocytic progenitor clusters resulted in the frequent misplacement of these respective cells into other cell populations based with similar transcriptomes (e.g., monocytic, granulocytic, MDP), whereas removal of HSCP-1 and HSCP-2 resulted in poor performance, as these cell populations are highly similar.

While alignment accuracy for cellHarmony is generally high, to further evaluate its performance in relation to similar methods we compared to the recently described scmap algorithm. scmap employs a related strategy for cell alignment with multiple algorithms used for assignment (Pearson, Spearman, cosine) as well as its own built-in feature selection method (**Fig. 2G**). When applied to our BM dataset, scmap resulted in a low Cohens kappa of 0.45 of cells with 41% of all cells excluded (not aligned) with the default thresholds. However, when the scmap analysis was restricted ICGS genes, Cohen's kappa was comparable to cellHarmony (0.76 with only 11.6% of the cells excluded) (**Fig. 2H)**.

Batch Effects

Although cellHarmony does not explicitly correct for batch effects, it is designed to classify cell types independent of such effects. This is accomplished by defining population-specific genes and cell-cluster references (e.g., ICGS on a single dataset) that are not conflated with donor or

batch effects. To evaluate, we examined the classification of scRNA-Seq from the Human Cell Atlas project, in which multiple donors and multiple captures for each donor were generated. Using data from human bone marrow donors (n=8), first we performed ICGS on each individual donor and then combined the resultant MarkerFinder results using the cellHarmony merge function (Hay et al., 2018). This merge function produces a non-redundant combined reference from the cellular medoids, composed of distinct MarkerFinder outputs. When such datasets are independently analyzed with ICGS, we assume that cell-type differences as opposed to donor effects will be identified. Although no clear differences could be identified from the different batches for the same donor, donor-skewed population predictions were evident using the software Seurat when blinded to donor composition (**Fig. 3A**). Following the identification of putative cell-populations with the cellHarmony merge function, projected cell-types showed a similar donor-skewed distribution in this data, with specific donor effects evident for multiple cell populations, notably, Neutrophils, Immature Neutrophils and Dendritic cells. (**Fig. 3B**). However, re-analysis within Seurat of the cellHarmony merged reference genes minimized these donor effects and allowed for largely coherent unbiased identification of distinct hematopoietic cell-types (**Fig. 3C,D**).

Differential Expression Analysis of Aligned Cell Populations

Numerous algorithms for scRNA-Seq differential expression analysis have been developed and compared to bulk RNA-Seq methods (Jaakkola et al., 2017). cellHarmony applies an empirical Bayes (eBayes) moderated t-test method, designed for bulk RNA-Seq and microarray studies, to identify genes meeting a user-defined fold and p-value cutoff. Comparison of scRNA-Seq differentially expressed genes for T-cells versus B-cells with eBayes, SCDE and MAST, relative to bulk RNA-Seq for these populations, found almost no differences in the overall accuracy of predictions from these methods (<2% difference) (**Fig. S2**).

Comparison of Related Cells Across Datasets

We next examined previously evaluated datasets in which cell-population predictions and/or cell-type specific gene expression changes were described by the original authors. The hematopoietic system again provides a useful proof of concept for evaluation, as associated cell populations and disease association changes have been largely resolved. Recently, we demonstrated the existence of distinct mixed-lineage cell populations in mouse bone marrow with multilineage potential (Olsson et al., 2016). These populations include a relatively frequent group of cells with myeloid, erythroid and megakaryocyte coincident expression (Multi-Lin), progenitors

in a metastable uncommitted state (GG1) and bi-potential intermediates that produce specified monocytic or granulocytic progenitors (IG2). Indeed, other research groups have isolated cells with similar predicted potential, such as myeloid-restricted pre–granulocyte-macrophage progenitors (aka pre-GM) (Drissen et al., 2016). CellHarmony of GG1 and IG2 cells into the published reference ICGS expression results aligns these most specifically to distinct progenitor populations, in agreement with previously published predictions (**Fig. 4A**). Specifically, IG2 cells are localized to a subset of the monocytic progenitor cluster with monocytic/granulocytic priming. While nearly half of all GG1 cells aligned to these same cells, GG1s (unlike IG2s) are frequently localized to Multi-Lin cells (**Fig. 4B**). These predictions align with the experimentally determined colony forming potential of these distinct cells. Although produced from an independently laboratory, Pre-GM's aligned to Multi-Lins, monocytic/dendritic progenitors and megakaryocytes, as predicted by the original authors (**Fig. 4C**). In addition to classifying subsets of cells, cellHarmony can organize a large dataset with comparable cell populations against established references. To demonstrate, we applied ICGS centroids from Fig. 4C to thousands of hematopoietic progenitors that were sequenced to a relatively shallow depth using the inDrops platform (GSM2388072) (Tusi et al., 2018). This analysis identified comparable cell populations with similar transcriptomes for the large majority of cells in the dataset (3,870 out of 4,535 cells, rho>0.3) (**Fig. 4D**).

In the above examples, all hematopoietic progenitors are present at steady-state. To assess progenitor frequencies in disease, we compared genetically perturbed hematopoietic progenitors derived from a mouse model of human Acute Myeloid Leukemia (AML) carrying both Flt3-ITD and Dnmt3a variants (Meyer et al., 2016). Splenic c-kit positive progenitors from these animals were aligned to the same bone marrow progenitor reference from the same strain of animals, collected from the same laboratory using the same scRNA-Seq methodology. Aligned cell-populations roughly matched the independently obtained AML populations reported from the original study (e.g., HSCP-like1, MDP-like and neutrophil-like) (**Fig. 4E**). A novel intriguing observation is that AML progenitors frequently localize with IG2 cells. These data suggest that a genetically defined subset of AMLs may derive from a short-lived cellular intermediate with bipotential monocytic and granulocytic potential (IG2). Differentially expressed genes calculated from the comparison of the AML cells to their cellHarmony matched populations reveals distinct patterns of up and downregulation in different cell populations (**Fig. 4F**). Direct comparison of the genes within these populations revealed both lineage-restricted differences as well as common gene expression changes, linked to experimentally validated findings (e.g., *Il18r1*) (**Fig. 4G**).

Dissection of Disease Cell-States from Clinical Samples

Given our findings in mouse, we applied the same workflow to previously generated human patient specimens with AML. To minimize potential batch and donor effects for differential analyses, we selected scRNA-Seq from a pre-transplantation AML bone marrow biopsy relative to a post-transplantation biopsy on the same patient, although the donor genetics will differ from the recipient. In addition to aligning the samples against each other (pre versus post), we directly aligned both to 35 bone marrow hematopoietic cell populations derived from an independent analysis of 8 healthy donors collected from the Human Cell Atlas (HCA) project (see Methods) (Hay et al., 2018). Alignment of these data to the HCA references identified variable numbers of samples associated with distinct progenitor and committed cell populations (**Fig. 5A**). These data suggest that this patient exhibits a wide-spread increase in the Erythroblast compartment prior to transplant (~5 fold increase), with a broad decrease in early progenitor populations (e.g., HSC, megakaryocytic, erythroid, granulocytic, early-erythroid) and committed cell populations (e.g., naïve T-cell, platelets, Eosinophil), compared to post-transplantation. Consistent with this observation, this patient was diagnosed with erythro-leukemia, which is characterized by proliferation of erythroblastic precursors. Importantly, these observations were reflected in the direct comparison of pre and post samples (**Fig. S3A**). Although variable numbers of cells aligned to these distinct populations, differential expression between the cellHarmony aligned populations revealed differences in the magnitude and direction of changes in different cell-types (**Fig. 5B**). Interestingly, the most divergent gene expression differences are found between Erythroblasts and CD34+ early erythroid progenitors, which showed preference for downregulation and upregulation in pre- vs. post-transplant, respectively. Global comparison of prior gene signatures among these sets of differentially expressed genes suggests regulation by distinct cancer gene networks in different hematopoietic subsets (**Fig. S3B**). A selective examination of differential genes from Erythroblasts finds previously observed genes up-regulated or down-regulated in leuekmia and clonal variation in the Erythroblast compartment prior to transplantation (Lazarini et al., 2016; Maitta et al., 2011; Martens et al., 2010).

Discussion

The accurate classification of single-cell profiles between datasets is a requirement for future efforts to reproducibly understand cellular diversity from scRNA-Seq datasets. Although unsupervised analysis approaches, such ICGS, can readily define cell populations in a *de novo* manner, such predictions will remain highly speculative without comparison to prior knowledge. Here, we demonstrate the ability of our new software, cellHarmony, to accurately classify cells

from data generated from the same and independent laboratories within the entirety of a reference dataset by performing cell-to-cell mapping. As we demonstrated in the hematopoietic system, this approach is able to identify rare or meta-stable cell populations (e.g., Multi-Lin's, IG2) across datasets, even when the original authors were unable to find such on their own. For alignment, this approach is able to overcome differences occurring due to batch or technology effects through the use of a non-batch conflated reference with reliable population-specific marker genes. For differential expression analysis, this approach uses the propagated cell-cluster labels to assess discrete differences between two reasonably comparable datasets, in which donor or batch effects are not found or have been corrected for outside of cellHarmony.

We should note, the direct mapping of cells between datasets has several potential advantages and disadvantages. The mapping of cells to obtain optimal matches allows for the detection of extremely rare sets of cells, which may represent unique biologically informative cell states (e.g., transitional cell populations), technical artifacts (multiplets) or simply mis-clustered cells. In the case of transitional cell populations, such as IG2, these could only effectively be aligned to a previously published reference given the entirety of the single-cell dataset. Furthermore, when mapping data across technological platforms, correlation to reference centroids will be more stable than to individual cells which are prone to high dropout gene expression. While cellHarmony suggests molecular differences that presumably are not dependent on batch effects or secondary genetic differences in the compared perturbation dataset, we recommend independent replication of the experiment to confirm such differences. Nonetheless, statistical confidence in the molecular differences underlying two compared datasets will be dependent on the number of cells aligning to a given population. With increasing use and decreasing expense of scRNA-Seq technologies, we anticipate approaches such as cellHarmony to become necessary to derive higher order insights into the investigation of pharmacological and disease heterogeneity.

ACKNOWLEDGMENTS

We thank Dr. Harinder Singh for his critical discussions related to this method. This work was supported by Cincinnati Children's Hospital Research Foundation and funding from the National Institutes of Health R01CA196658, R01HL122661 and R21AI35595 (HLG).

AUTHOR CONTRIBUTIONS

The manuscript was written by NS and ED. The method was conceived by NS and HLG. Computational analyses and algorithm evaluations were performed by ED, KF, SH and NS.

FIGURE LEGENDS

Figure 1. The cellHarmony workflow for matching and comparing single-cell transcriptomes. A) Illustration of the cellHarmony workflow when comparing a query (i.e. disease) to a reference (i.e. healthy) single-cell dataset. B-C) Distinct outputs of the cellHarmony workflow to assess (B) cell frequency differences and (C) cell-specific transcriptomic perturbations between compared samples.

Figure 2. cellHarmony accurately classifies cells from distinct and granular single-cell populations. A-B) Evaluation of mouse bone marrow hematopoietic progenitor scRNA-Seq. A) The heatmap is the primary ICGS output used for cellHarmony (tab-delimited text file), in which 10% of all cells are removed and tested against the remaining cells over 100 independent rounds. B) Inter-rater agreement assessed using Cohen's kappa (unweighted) at various cellHarmony correlation cutoffs (left panel) and the percentage of cells placed at each cutoff (middle panel). The right panel indicates the combined performance metric (kappa and placement). C-D) Evaluation of human melanoma patient scRNA-Seq (putative doublets removed) by ICGS using the same strategy outlined in panel A and B. E) Frequency of cell population mis-assignments by cellHarmony for the bone marrow dataset. The original source cell clusters are indicated for the tested cells, followed by the assigned, misplaced population. F) To assess the specificity of cell placements using distinct correlation cutoffs, all cells from each indicated cell-cluster were removed and aligned back to the remaining cellHarmony clusters along a random set of selected test cells (10%). The differing inter-rater agreement is indicated for each panel with removal of the indicated cell cluster. Poor inter-rater agreement indicates that cellHarmony incorrectly classified the removed cluster cells into another cluster for that indicated correlation threshold. G) The top 500 variable genes identified from scmap using default options, based on gene expression intensity and dropout. H) Inter-rater agreement with Cohen's Kappa (unweighted) for scmap using its own variable genes or those explicitly provided from ICGS.

Figure 3. Cell-population specific marker genes reduce batch effect bias. Demonstration of the improved use of pre-selected marker genes for cellHarmony alignment with different donor samples. Cell populations and associated marker genes were independently derived using donor-specific ICGS analyses followed by cellHarmonyMerge analysis for all eight HCA donor samples (Methods). A-B) Seurat unsupervised analysis applied to all genes with putative cell populations

visualized for three combined bone marrow donor samples. A) Visualization of independently derived cell-type labels indicates significant batch effects as compared to B) visualization of donors on the t-SNE. C-D) Seurat analysis repeated using only the cellHarmonyMerge workflow population-specific genes.

Figure 4. Identification and characterization of transitional cell populations in perturbed and unperturbed hematopoiesis with cellHarmony. A) cellHarmony heatmap generated when aligning previously isolated bipotential hematopoietic progenitors (IG2) to ICGS results from all captured normal mouse bone marrow hematopoietic progenitors. A minimum Pearson correlation cutoff of 0.6 was applied. B-C) This analysis was repeated for mixed multi-potent cellular intermediates (GG1 and Pre-GM) obtained from two laboratories. D) cellHarmony classification of over 4,500 mouse bone marrow progenitors using the inDrops platform against the same Fluidigm bone marrow cell population reference with IG2 cells included. This analysis was performed using the centroid correlation option (--referenceType centroid) with a correlation cutoff of 0.3. We these options, 664 cells were excluded due insufficient centroid matching. E) cellHarmony alignment of prior profiled AML c-kit-positive cells from the mouse spleen relative to the same Fluidigm reference bone marrow ICGS results. F) The number of cellHarmony differentially up- and down-regulated expressed genes associated with each cell-population comparison (AML versus wild-type) from cellHarmony. G) cellHarmony produced sorted heatmap (MarkerFinder algorithm) of differentially expressed genes in AML versus normal matched cell populations. Genes in red indicate those experimentally verified in the original study using either bulk RNA-Seq or flow cytometry.

Figure 5. Changes in cell population frequency and gene expression in AML over-time. Analysis of bone marrow mononuclear cells obtained by scRNA-Seq (10x Genomics) from a single-patient (AML027) prior to and after bone marrow transplantation. A-B) cellHarmony classification of cells from AML027 A) post-transplantation and B) prior to transplantation. C) Number of patient cells assigned to prior annotated bone marrow cell populations (Figure 3) by cellHarmony. D) Magnitude of differential gene expression changes in post-transplantation associated cell populations versus pre-transplantation. E) Variation in Erythroblast gene expression among individual cells in the pre- and post-transplantation AML samples for cellHarmony differentially expressed genes. Previously defined leukemia up-or down-regulated genes are highlighted.

Figure S1. Evaluation of cellHarmony from Melanoma scRNA-Seq. A) Inter-rater agreement for Melanoma scRNA-Seq without putative doublet removal, assessed using Cohen's kappa (unweighted) at various cellHarmony correlation cutoffs (left panel) and the percentage of cells placed at each cutoff (middle panel). The right panel indicates the combined performance metric (kappa and placement). B) Frequency of cell population mis-assignments by cellHarmony in Melanoma with doublet removal included. The original source cell clusters are indicated for the tested cells, followed by the assigned, misplaced population. C) The differing inter-rater agreement is indicated for each panel with removal of the indicated Melanoma cell cluster. Poor inter-rater agreement indicates that cellHarmony incorrectly classified the removed cluster cells into another cluster for that indicated correlation threshold.

Figure S2. Suitability of different algorithms for scRNA-Seq differential expression relative to bulk RNA-Seq. Venn diagrams are shown comparing reported differentially expressed genes (DEGs) from T-cells and B-cells profiled either by bulk RNA-Seq or scRNA-Seq. Three algorithms for differential gene expression analysis (eBayes, SCDE and MAST) were applied for scRNA-Seq.

Figure S3. Comparison of human bone marrow mononuclear cells (BMMCs) prior to and after transplantation. A) cellHarmony alignment of pre-transplantation BMMCs to post-transplantation ICGS marker gene results (left) and for pre-transplantation alone (right). Gene set enrichment results in AltAnalyze are shown on the left of each heatmap gene-cluster against marker genes corresponding to 35 cell populations identified from an independent analysis of bone marrow hemopoietic cell populations (Methods). B) Comparison of gene-set enrichment results using the software GO-Elite for MSigDB signature gene sets for cellHarmony population associated differences (pre- versus post-transplantation, up- and down-regulated genes).

References

Butler, A., and Satija, R. (2017). Integrated analysis of single cell transcriptomic data across conditions, technologies, and species. bioRxiv.

DePasquale, E.A.K., Schnell, D.J., Valiente, I., Blaxall, B.C., Grimes, H.L., Singh, H., and Salomonis, N. (2018). DoubletDecon: Cell-State Aware Removal of Single-Cell RNA-Seq Doublets. bioRxiv.

Drissen, R., Buza-Vidas, N., Woll, P., Thongjuea, S., Gambardella, A., Giustacchini, A., Mancini, E., Zriwil, A., Lutteropp, M., Grover, A.*, et al.* (2016). Distinct myeloid progenitor-differentiation pathways identified through single-cell RNA sequencing. Nature immunology *17*, 666-676.

Haghverdi, L., Lun, A.T.L., Morgan, M.D., and Marioni, J.C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nature biotechnology *36*, 421-427.

Hay, S.B., Ferchen, K., Chetal, K., Grimes, H.L., and Salomonis, N. (2018). The Human Cell Atlas bone marrow single-cell interactive web portal. Exp Hematol.

Jaakkola, M.K., Seyednasrollah, F., Mehmood, A., and Elo, L.L. (2017). Comparison of methods to detect differentially expressed genes between single-cell populations. Briefings in bioinformatics *18*, 735-743.

Kiselev, V.Y., Yiu, A., and Hemberg, M. (2018). scmap: projection of single-cell RNA-seq data across data sets. Nature methods *15*, 359-362.

Lazarini, M., Machado-Neto, J.A., Duarte, A.D., Pericole, F.V., Vieira, K.P., Niemann, F.S., Alvarez, M., Traina, F., and Saad, S.T. (2016). BNIP3L in myelodysplastic syndromes and acute myeloid leukemia: impact on disease outcome and cellular response to decitabine. Haematologica *101*, e445-e448.

Magella, B., Adam, M., Potter, A.S., Venkatasubramanian, M., Chetal, K., Hay, S.B., Salomonis, N., and Potter, S.S. (2017). Cross-platform single cell analysis of kidney development shows stromal cells express Gdnf. Developmental biology.

Maitta, R.W., Wolgast, L.R., Wang, Q., Zhang, H., Bhattacharyya, P., Gong, J.Z., Sunkara, J., Albanese, J.M., Pizzolo, J.G., Cannizzaro, L.A.*, et al.* (2011). Alpha- and beta-synucleins are new diagnostic tools for acute erythroid leukemia and acute megakaryoblastic leukemia. Am J Hematol *86*, 230-234.

Martens, J.H., Brinkman, A.B., Simmer, F., Francoijs, K.J., Nebbioso, A., Ferrara, F., Altucci, L., and Stunnenberg, H.G. (2010). PML-RARalpha/RXR Alters the Epigenetic Landscape in Acute Promyelocytic Leukemia. Cancer cell *17*, 173-185.

Meyer, S.E., Qin, T., Muench, D.E., Masuda, K., Venkatasubramanian, M., Orr, E., Suarez, L., Gore, S.D., Delwel, R., Paietta, E.*, et al.* (2016). DNMT3A Haploinsufficiency Transforms FLT3ITD Myeloproliferative Disease into a Rapid, Spontaneous, and Fully Penetrant Acute Myeloid Leukemia. Cancer discovery *6*, 501-515.

Olsson, A., Venkatasubramanian, M., Chaudhri, V.K., Aronow, B.J., Salomonis, N., Singh, H., and Grimes, H.L. (2016). Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. Nature *537*, 698-702.

Prabhakaran, S., Azizi, E., Carr, A., and Pe'er, D. (2016). Dirichlet Process Mixture Model for Correcting Technical Variation in Single-Cell Gene Expression Data. JMLR Workshop Conf Proc *48*, 1070-1079.

Tusi, B.K., Wolock, S.L., Weinreb, C., Hwang, Y., Hidalgo, D., Zilionis, R., Waisman, A., Huh, J.R., Klein, A.M., and Socolovsky, M. (2018). Population snapshots predict early haematopoietic and erythroid hierarchies. Nature *555*, 54-60.

Figure 1

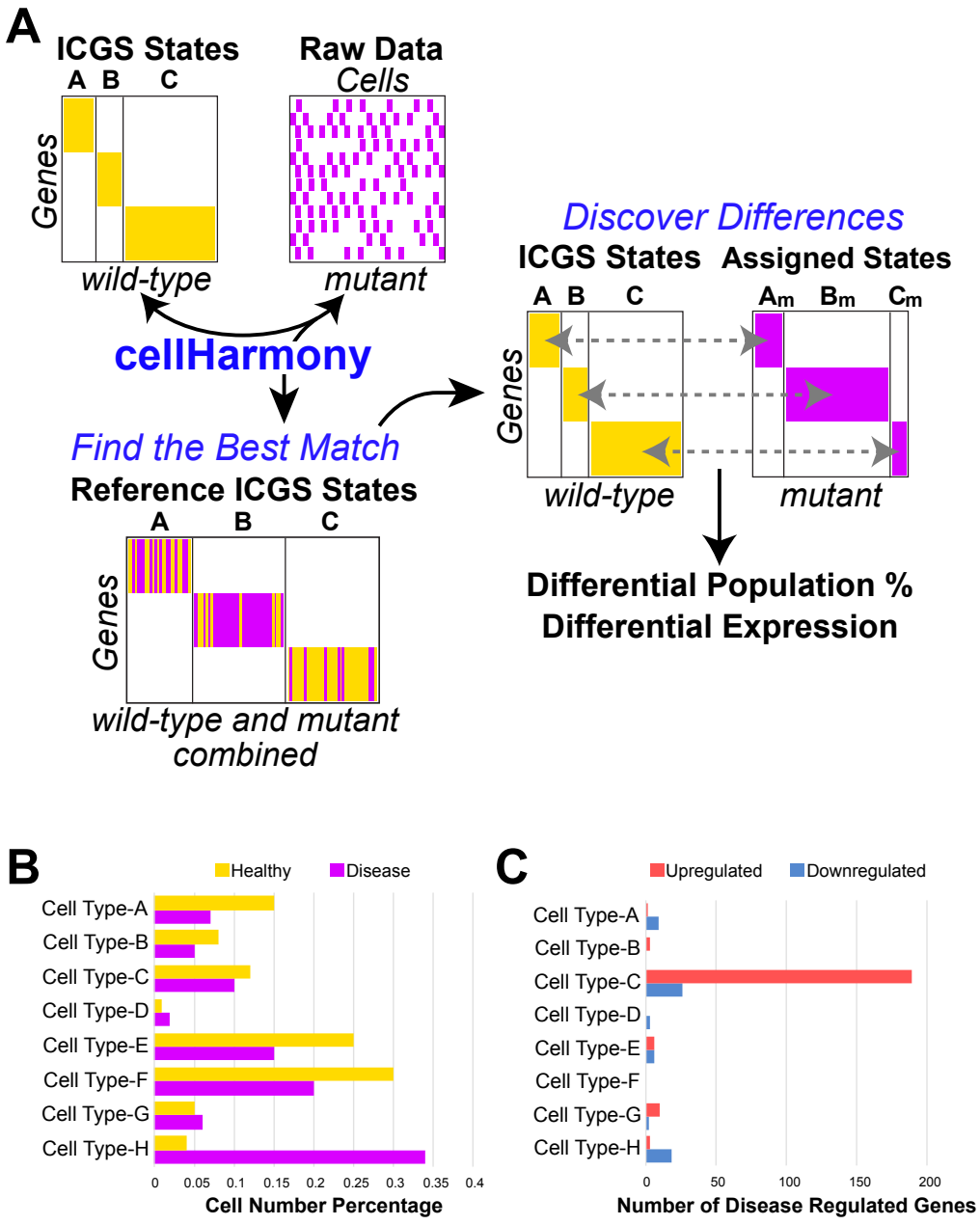

Figure 2

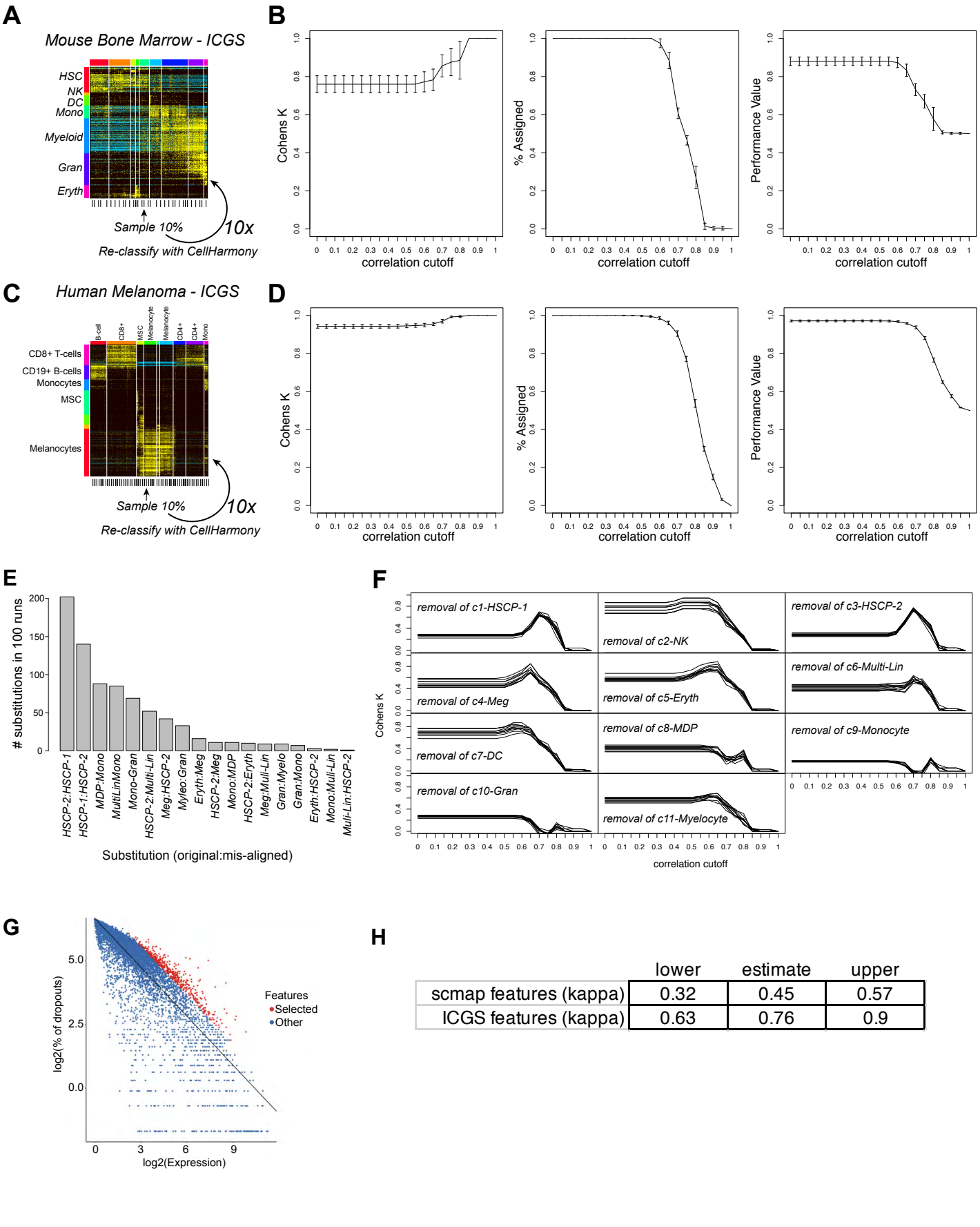

Figure 3

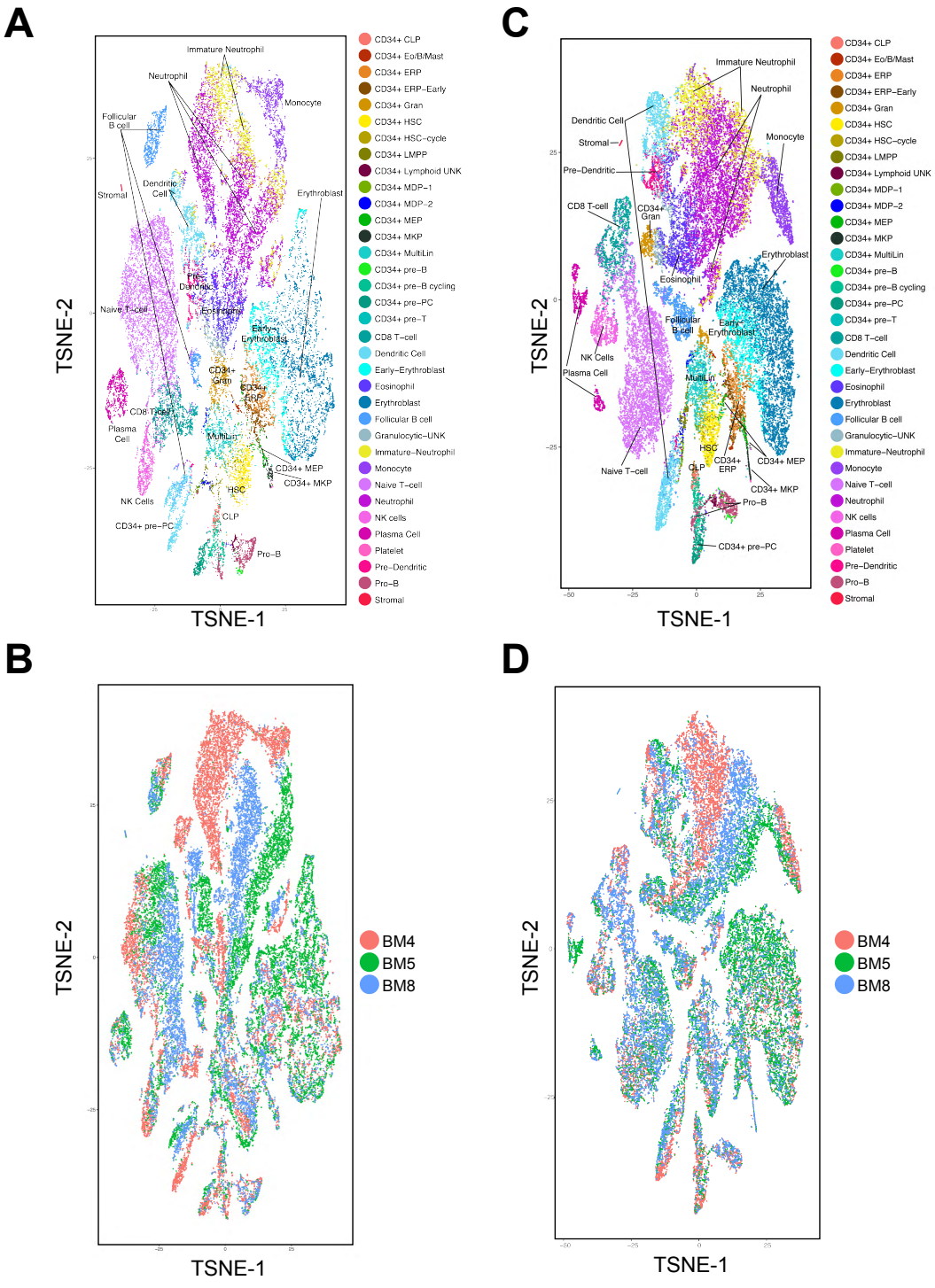

Figure 4

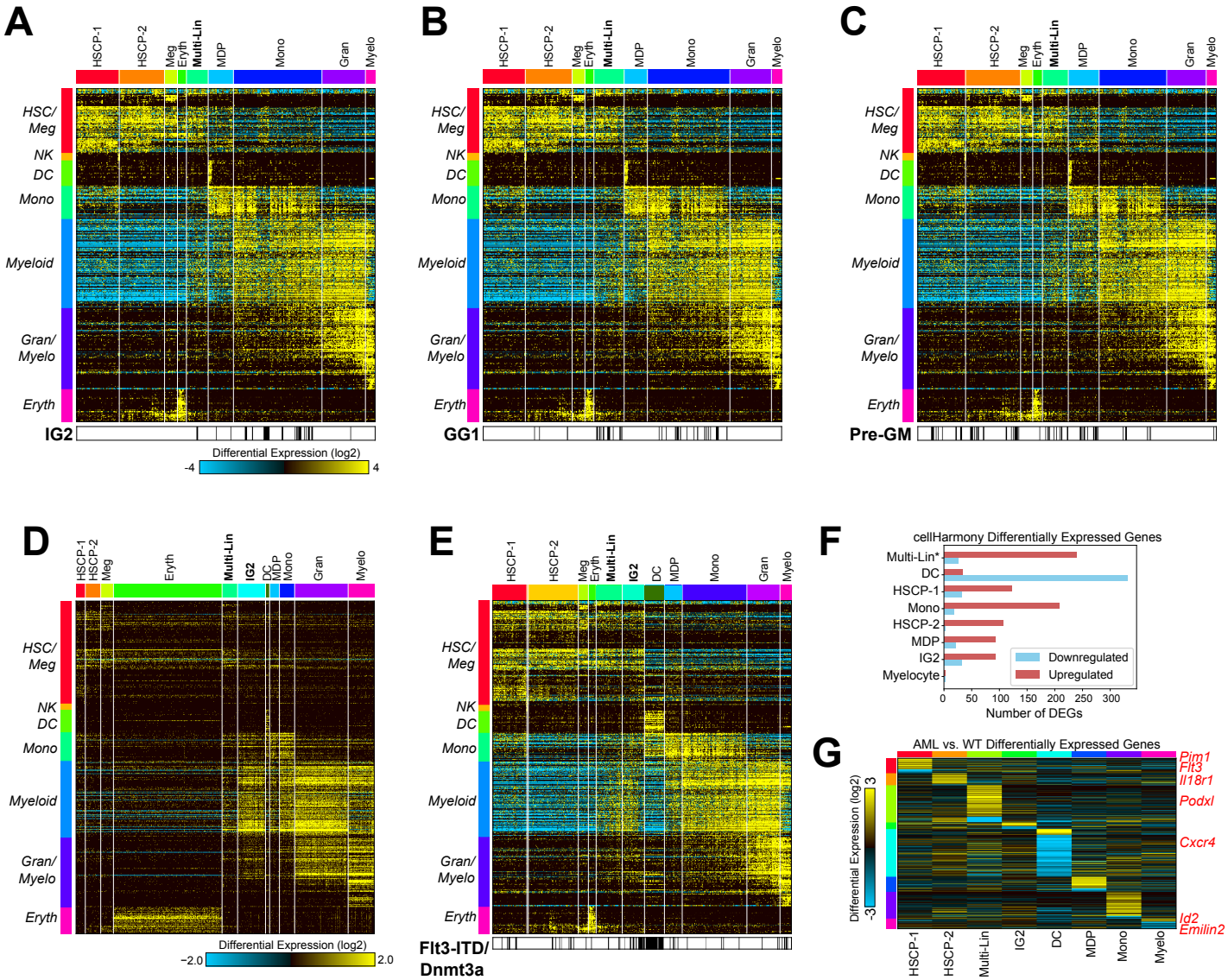

Figure 5

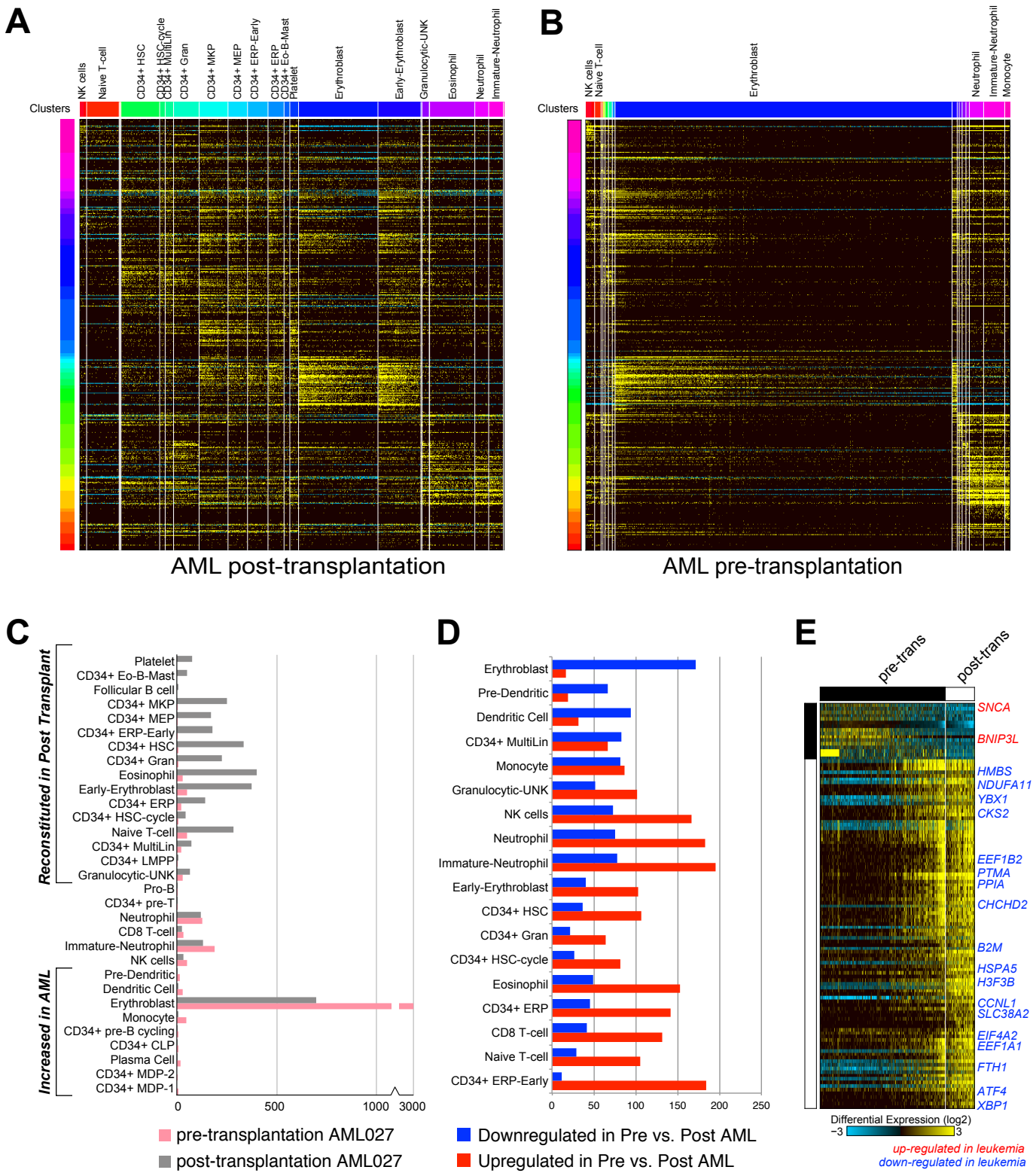

Figure S1

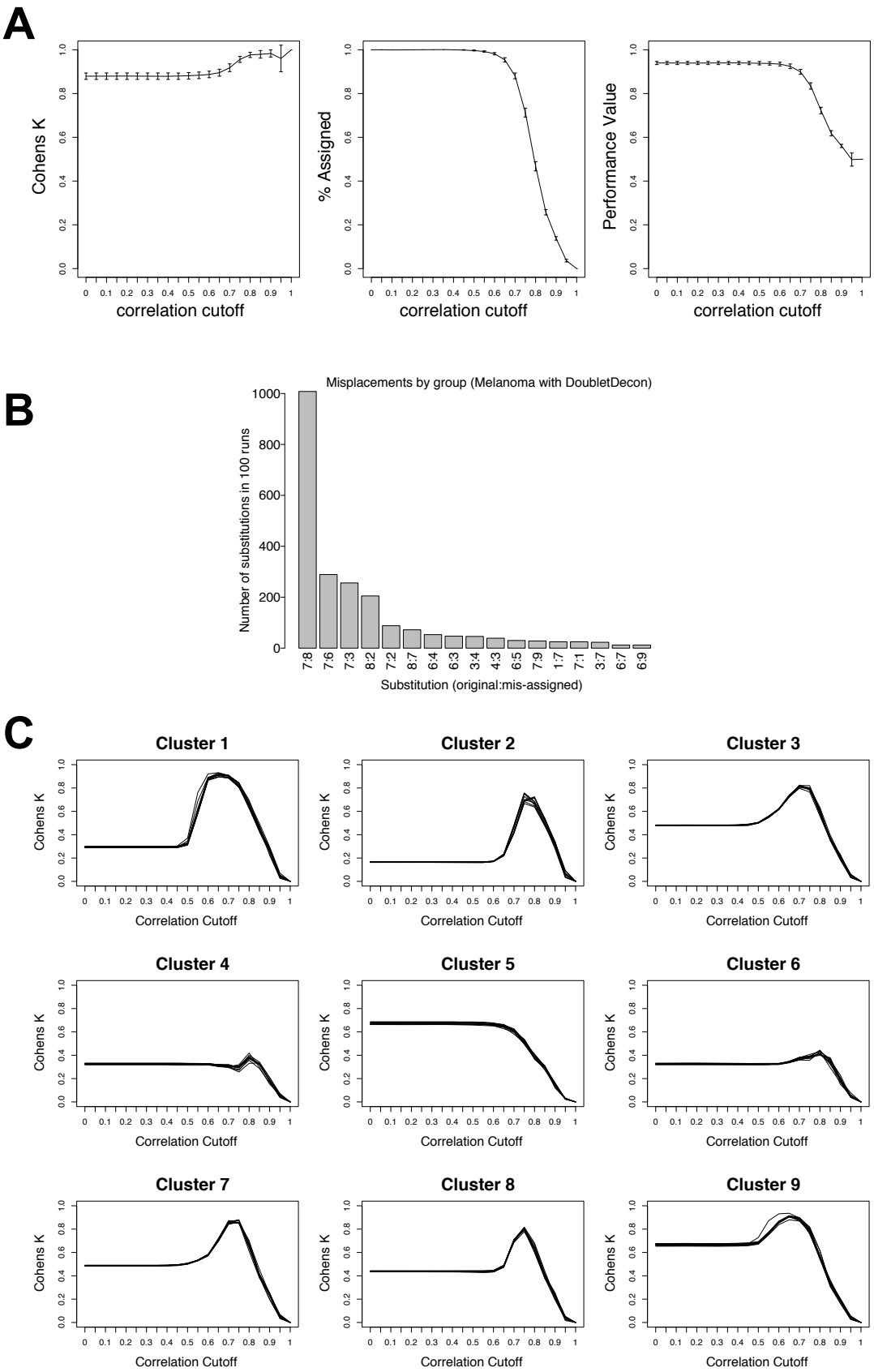

Figure S2

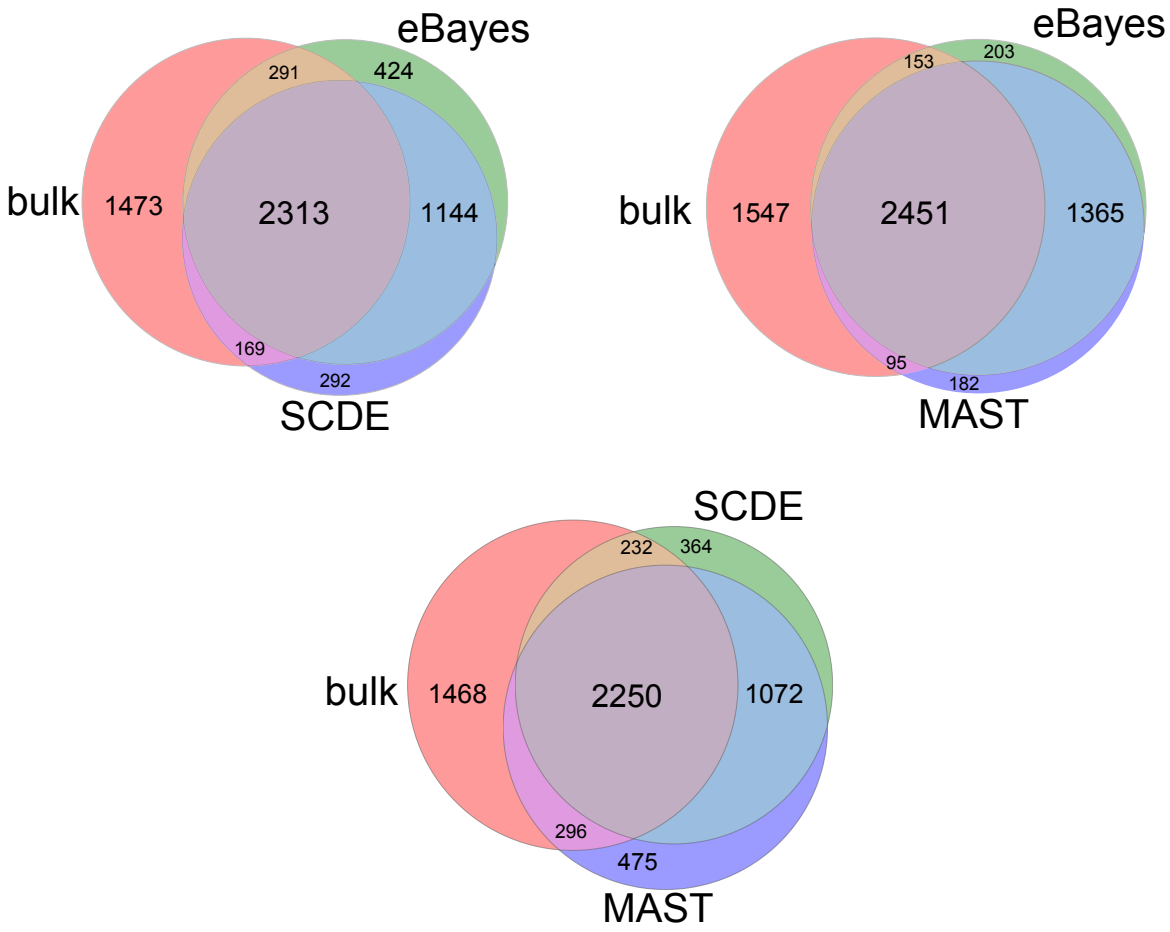

Figure S3