# *Supplementary Material*

# Batch effect correction methods for NASA GeneLab transcriptomic datasets

Lauren M. Sanders[1,2], Hamed Chok[3], Finsam Samson[4], Ana Uriarte Acuna[2,5], San-huei Lai Polo[2,5], Valery Boyko[2,6], Yi-Chun Chen[2,5], Marie Dinh[2,7], Samrawit Gebre[2], Jonathan M. Galazka[2], Sylvain V. Costes[2], Amanda M. Saravia-Butler[2,5*]

[1]Blue Marble Space Institute of Science, NASA Ames

[2]NASA Ames Research Center; Moffett Field, CA, 94035, USA

[3]GeneLab Multi-Omics Analysis Working Group

[4]Stanford University, Department of Computer Science; Stanford, CA, 94305, USA

[5]KBR, Space Biosciences Division, NASA Ames Research Center, Moffett Field, CA 94035, USA

[6]The Bionetics Corporation, NASA Ames Research Center, Moffett Field, CA 94035, USA

[7]Logyx, LLC, Mountain View, CA 94043, USA

***Correspondence:**
Corresponding Author
Amanda.M.Saravia-Butler@nasa.gov

## Supplementary Methods

We provide here additional details on our scoring method which identifies the optimal batch variable / correction method pair for a given dataset by geometrically probing the space of all allowable scoring functions (weighted combinations of evaluation criteria) to yield an aggregate volume-based scoring measure.

With variable/method pairs viewed as a point cloud in a vector space whose dimensions are the evaluation criteria, optimizing any given scoring function necessarily occurs at an extreme ("corner") point and thus any potential interior points are necessarily suboptimal. To obtain the point cloud's extreme points, the cloud's convex hull may be computed. The convex hull is defined as the set of halfspaces whose intersection enclose the point cloud and where each halfspace is defined via a hyperplane that exposes a convex hull face. To this end, an extreme point may be viewed as the intersection of at least d (non-degenerate) hyperplanes where d is the dimension of the data's vector space. Each hyperplane may be defined in terms of a

normal vector (taken in the positive direction following the ‘maximization’ convention). The set of normals associated with a set of hyperplanes collectively exposing a particular extreme point may be viewed as a set of generators whose linear combination in terms of any positive coefficients yields a scoring function that is maximized precisely at the underlying extreme point. Thus, it follows that any point in the cone generated by the normals in question yields a scoring function that is a maximizer for the associated extreme point.

Since only scoring functions with positive coefficients summing up to 1 are of interest, the space comprised of all maximizing scoring functions associated with a particular extreme point happens to be the intersection of the normals-generated cone and the standard simplex. The volume of such an intersection provides a measure of how “popular” a candidate method is (i.e., the “span” of scoring functions that are optimized at the extreme point representing said candidate method).

Such an intersection is said to form a category of scoring functions as they are all equivalent in terms of which candidate method they optimize. A direct category volume comparison reveals which candidate method is more “popular” in the “universe” of all possible scoring functions (i.e., d-simplex). In addition to evaluating popularity (i.e., a category’s volume), constituent scoring functions for any given category may be computed. Constituents are defined such that their mixtures span their associated category. Constituents from all categories may be useful when, for instance, an arbitrary scoring function (e.g., input by a user) is given and its category of equivalent scoring functions is to be determined. Such a determination provides an answer-product to the user in terms of how categorically popular their scoring preference is.

One final characterization is in terms of category centroids. Centroids are geometric centers which can be used to provide an intuitive way to compare and contrast scoring categories (e.g., Supplementary Table 3).

## Supplementary Tables

**Supplementary Table 1: Full sample-level metadata for the combined dataset (attached).**

**Supplementary Table 2. Comparison of DEGs in FLT vs. GC groups within each dataset before and after correction.** The "Uncorrected" row shows the number of differentially expressed genes (DEGs) in each dataset before correction. For each batch variable and correction method combination, the number of DEGs that match the original uncorrected DEGs are shown outside of the parentheses, while the number of DEGs that were identified only after correction are shown within parentheses.

| | | GLDS 47 | GLDS 48_I | GLDS 48_C | GLDS 137 | GLDS 168 | GLDS 173 | GLDS 242 | GLDS 245_LAR | GLDS 245_ISST |
|---|---|---|---|---|---|---|---|---|---|---|
| | Uncorrected | 14 | 63 | 197 | 3 | 1401 | 520 | 321 | 39 | 539 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| LibPrep as Batch | MBatch_AN | 4 (84) | 6 (277) | 120 (345) | 1 (33) | 147 (84) | 95 (85) | 143 (274) | 12 (37) | 317 (276) |
| | MBatch_EB | 3 (60) | 23 (159) | 95 (170) | 0 (8) | 305 (35) | 105 (49) | 173 (204) | 24 (45) | 414 (208) |
| | MBatch_MP | 5 (394) | 28 (914) | 150 (1823) | 1 (358) | 528 (2351) | 175 (715) | 125 (1088) | 32 (1330) | 379 (3062) |
| Mission as Batch | MBatch_AN | 2 (67) | 9 (72) | 19 (52) | 0 (20) | 15 (18) | 21 (81) | 12 (24) | 6 (33) | 180 (108) |
| | MBatch_EB | 4 (77) | 15 (87) | 30 (56) | 0 (5) | 23 (19) | 5 (5) | 37 (80) | 18 (27) | 299 (172) |
| | MBatch_MP | 12 (651) | 31 (538) | 108 (876) | 0 (36) | 585 (2861) | 484 (3271) | 218 (1528) | 28 (492) | 381 (2258) |

## Supplementary Table 3. Scoring coefficients for each evaluation criteria (category centroids) for each batch variable / correction method pair.

| | BatchQC skew | BatchQC kurt | PCA batch | PCA cond | DSC | LFC | DGE |
|---|---|---|---|---|---|---|---|
| **ComBatseq_libprep** | 0.145840644 | 0.206307139 | 0.18924 | 0.11127 | 0.128051 | 0.035856 | 0.183435 |
| **ComBat_libprep** | 0.116892382 | 0.110018744 | 0.163014 | 0.110364 | 0.152126 | 0.228696 | 0.118889 |
| **AN_libprep** | 0.138541655 | 0.156682548 | 0.060553 | 0.285797 | 0.031312 | 0.253622 | 0.073492 |
| **EB_libprep** | 0.139135134 | 0.128192181 | 0.178758 | 0.089391 | 0.018236 | 0.249198 | 0.197089 |
| **MP_libprep** | null | null | null | null | null | null | null |
| **ComBatseq_mission** | 0.326306213 | 0.31628422 | 0.050109 | 0.122816 | 0.034871 | 0.087038 | 0.062576 |
| **ComBat_mission** | 0.194474145 | 0.152704366 | 0.104779 | 0.133019 | 0.27432 | 0.075897 | 0.064807 |
| **AN_mission** | null | null | null | null | null | null | null |
| **EB_mission** | 0.25305187 | 0.178994381 | 0.327088 | 0.11433 | 0.011308 | 0.077625 | 0.037603 |
| **MP_mission** | 0.082274038 | 0.081790738 | 0.077758 | 0.308942 | 0.084032 | 0.103843 | 0.26136 |