Bundle correction plan for ICA Preview Release dataset

This document gives an overview of plans to correct the Immune Cell Atlas (ICA) Preview Release bundles.

Current bundling and ICA bundle structure

The “current” ICA bundle structure refers to the bundle structure that was implemented for the ICA dataset in the HCA Preview Release that occured on April 2, 2018.

Currently, an ICA bundle is defined in the metadata spreadsheet by the uniqueness of the value in the process_id column in the sequence_file tab. Simply, all files associated with the same sequencing process_id will be put into the same ICA bundle with all upstream biomaterials and processes/protocols included.

Issues with current bundling and ICA bundle structure

In the ICA Preview Release, only 4 unique process_ids were used in the spreadsheet, one ID for each of four Illumina sequencing flow cells (8 lanes per flow cell) used in the study. This resulted in the generation of 4 ICA bundles, each containing multiple biomaterials/processes/protocols/files.

The ICA bundle structure results in not enough granularity for the Data Portal. For example, there are multiple donors per ICA bundle in the current schema, so searching for data related to a single donor returns an ICA bundle with data from many donors in it.

The ICA bundle structure also results in the wrong granularity for Secondary Analysis Pipelines, which are expecting data for a single “sample” - in this case for a single cell suspension - per ICA bundle. In the current structure, one ICA bundle contains 192 data files (186 for an ICA bundle that is missing 1 biomaterial) representing 3 10x files for 2 technical replicates (each replicate run on 1 lane) for 8 pooled cell suspensions across 4 pairs of lanes on the flow cell (3 x 2 x 8 x 4 = 192 files). What the analysis pipelines need is a single ICA bundle with the 3 10x files, demultiplexed by technical replicates and biomaterial (in this case, cell suspension).

Proposed new bundling and ICA bundle structure

Owing to recent ambiguity over ICA bundle definition, we now propose a concrete definition of an ICA bundle as:

*The files and metadata generated from the smallest **biomaterial** entity in the metadata hierarchy including separation of technical replicates.***

Under this definition, an ICA bundle structure will be technique-dependent. For example, for 10x experiments, a bundle will be composed of 3 files (I1, R1, R2) for the smallest biomaterial.

The current ICA bundles in the Production Data Store are incompatible with how the Data Portal and the Secondary Analysis Pipelines need to function. The Ingestion Service is therefore proposing to re-ingest the ICA metadata and data files with minimal changes to the metadata spreadsheet in order to generate DCP component-compatible ICA bundles.

This proposal does not contain changes or updates to ingestion code. The changes here are a quick fix in order to get correct ICA bundles in the Data Store as soon as possible, with more stable fixes using ingestion code forthcoming.

****NB**. This definition holds true for the other 2 datasets in the HCA Preview Release, and will hold true for all future bundles created by the Ingestion Service.

Changes needed to implement new ICA bundle structure

Changes needed to migrate the current ICA bundles to a compatible structure are only required at the level of the process_id linking column in the metadata spreadsheet that is uploaded to Ingest. There are no code or other metadata changes required at this time.

Spreadsheet

Changes are required in two tabs in the ICA metadata spreadsheet.

1. *sequencing_process* tab. The sequencing_process tab, which currently contains 4 sequencing processes each generating 1 bundle, will need to be expanded to have 254 unique sequencing processes. Each of these processes will be linked to 1 cell suspension/lane combination in the *sequence_file* tab to generating 1 bundle (for 254 bundles in total).
2. *sequence_file* tab. The sequence_file tab, which currently contains only the 4 unique sequencing process IDs in the process_id column, will need to be changed to contain the 254 unique sequencing process IDs, one ID for each triplicate of I1, R1, and R2 files associated with a single cell suspension/lane combination. Although there are 127 cell suspensions, each suspension was run on 2 lanes representing technical replicates. These technical replicates need to be treated separately by the secondary analysis pipelines and therefore need to be in their own bundle.

Ingest code

No changes are required to ingest code from what was used in Production for the Preview Release for bundle correction. This includes the importer, ingest-broker, and exporter.

Metadata

Besides expanding the sequencing process IDs in the two tabs indicated in “Spreadsheet” section, there are no additional changes are required to the actual metadata for bundle correction.

Submission

No changes are required to the actual process of uploading, validation, and submitting the ICA dataset with the updated spreadsheet, but the submission process will need to be repeated (in Production environment) from start to end to export new correct bundles to the Data Store.

Effect of changes on DCP components and preview site

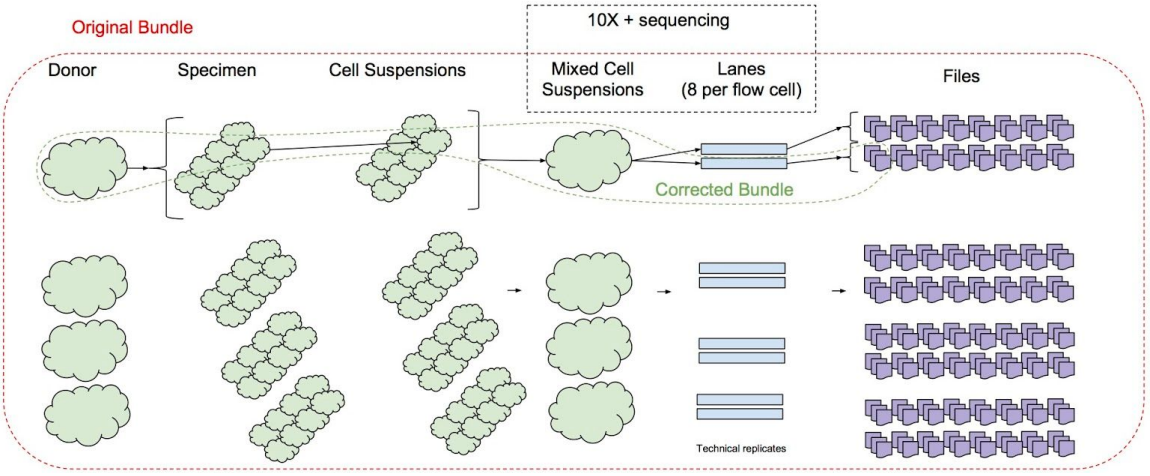
We are seeking clarification with Secondary Analysis Pipelines and the Data Portal on whether the proposed new bundling scheme for the ICA Preview dataset meets their needs. Some issues to take into account include:

1. Conveying on the Preview webpage that new bundles are being released.
2. Confirmation before re-releasing ICA dataset that the bundles are as expected for triggering secondary analyses, especially in terms of using technical replicates

We are also seeking information from the Data Store team about how to implement an update to the ICA Preview Dataset. Some issues to take into account include:

1. This change constitutes an “update” to the current bundles, although ICA will be going from 4 bundles to 254 bundles.
2. We will need to re-ingest ICA in the Production environment, so Ingest and Data Store need to be coordinated on this.
    a. https://github.com/HumanCellAtlas/data-store/issues/1207
3. The download scripts for the ICA dataset might change if they are based on UUIDs.
    a. https://github.com/HumanCellAtlas/dcp-preview/issues/33

Diagram of current (original) and proposed (corrected) ICA bundle



Current (original) ICA bundle:

- Contains all 192 data files generated on a single Illumina flow cell (across 8 lanes)
- Contains all data for 4 individual donors
- Contains all data from 4 pools of 8 cell suspensions (“Mixed Cell Suspensions”); each pool is split across 2 lanes (representing technical replicates)

Proposed (corrected) ICA bundle:

- Will contain 3 data files ONLY which trace back to 1 cell suspension
- Will be demultiplexed per cell suspension and therefore traced back to 1 specimen and 1 donor