The Scanpy Single Cell Sandbox: A Terra notebook to improve analytic access to single cell data

Amelia Weber Hall

PCL Meeting 2/18/2020

The single cell processing system

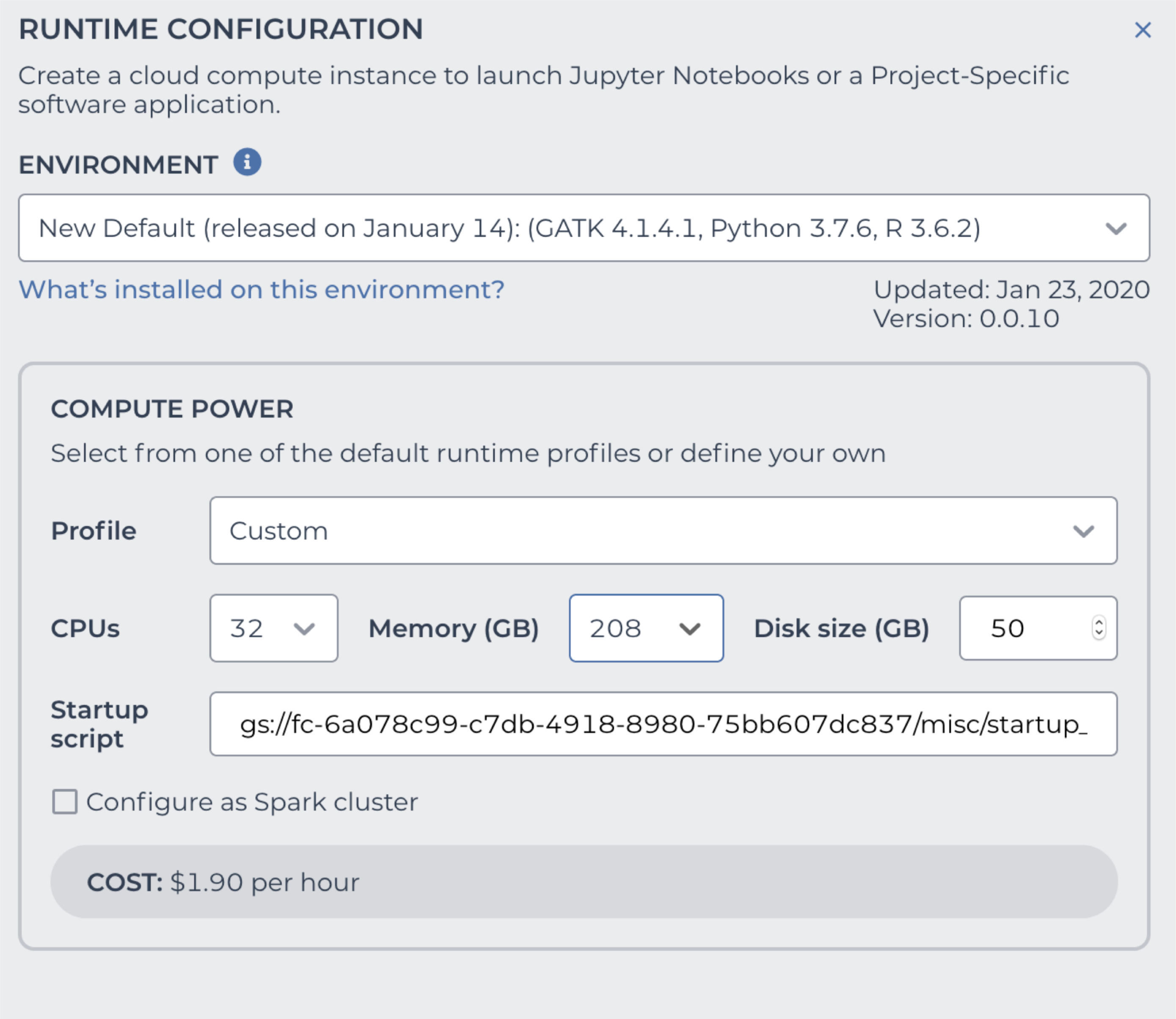
- The skillsets for computational biologists and bench biologists are distinct
    - Bench biologists generate the libraries and primary sequence data
        - (and troubleshoot platform issues with 10X, modify/optimize protocols for nuclear isolation/RNA amplification)
    - Computational biologists process raw sequence data into harmonized single cell data objects
        - (and identify marker genes for subpopulation clusters, use scVI to reduce background RNA influence on clustering, infer and correct for batch effects)
- However, bench biologists need to have relatively easy access to specific and custom analyses of these data
    - Both for science and for iterative bench biology reasons

How can we expand access to single cell data/ analyses?

- Problems:
  - Prem access for Bayer folks
    - Data sharing issues
  - The size of the single cell data objects
  - Analyst workload (esp for custom) plots and figures

- Solutions:
  - Use Terra for analysis
  - Use buckets to store objects centrally
    - So they can be copied on demand
  - Build a sandbox that handles most basic analyses for bench biologists
    - Justification: folks who can follow complex protocols can do the same in a computational setup (provided good documentation)

Notebook setup: H4C is huge



RUNTIME CONFIGURATION

Create a cloud compute instance to launch Jupyter Notebooks or a Project-Specific software application.

ENVIRONMENT

New Default (released on January 14): (GATK 4.1.4.1, Python 3.7.6, R 3.6.2)

What's installed on this environment?

Updated: Jan 23, 2020
Version: 0.0.10

COMPUTE POWER

Select from one of the default runtime profiles or define your own

Profile: Custom

CPUs: 32

Memory (GB): 208

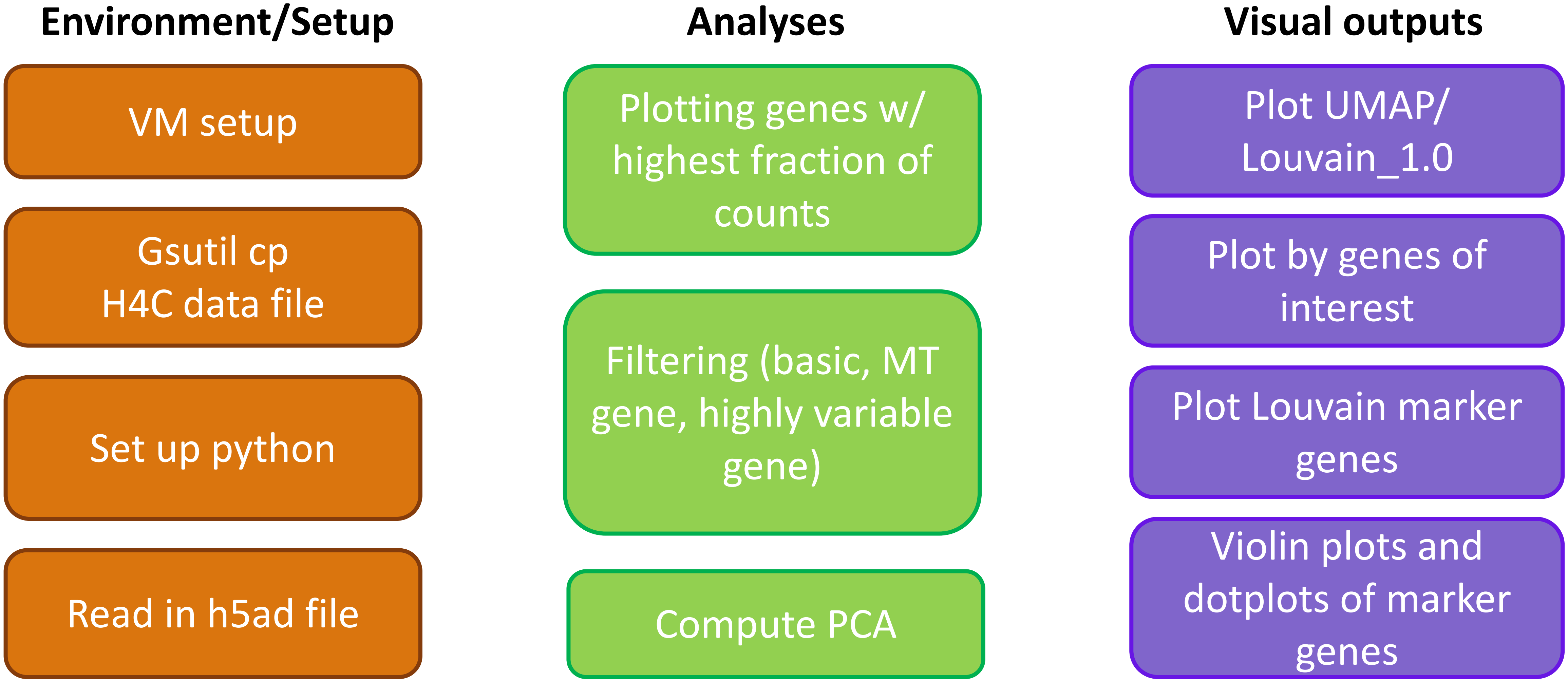
Disk size (GB): 50

Startup script: gs://fc-6a078c99-c7db-4918-8980-75bb607dc837/misc/startup_

□ Configure as Spark cluster

COST: $1.90 per hour

Workflow of the sandbox



Environment/Setup

- VM setup
- Gsutil cp H4C data file
- Set up python
- Read in h5ad file

Analyses

- Plotting genes w/ highest fraction of counts
- Filtering (basic, MT gene, highly variable gene)
- Compute PCA

Visual outputs

- Plot UMAP/ Louvain_1.0
- Plot by genes of interest
- Plot Louvain marker genes
- Violin plots and dotplots of marker genes

https://app.terra.bio/#workspaces/bayer-pcl-single-cell/single%20cell%20sandbox%20scanpy/notebooks/launch/custom_gene_list_visualization.ipynb?mode=edit

Future Plans/Unsolved Questions

- How much pre-plotting normalization/filtering is required/necessary for this dataset?
  - Different threshold recommendations for different purposes?
- Easy input and plotting for large numbers of genes?
- Subclustering of individual subgroups and clusters?
- Best ways to identify marker genes?
  - AUC calculation?
- Any other useful or highly desired features?