Libraries:

```
library(dplyr)
library(dtplyr)
library(tidyr)
library(data.table)
library(patchwork)
library(ggplot2)
DATA_DIR <- "/Users/chris/Desktop/DSC_291/Final_Project_Data"
setwd(DATA_DIR)
knitr::opts_knit$set(root.dir = DATA_DIR)
```

Annotations:

```
# Use PanglaoDB for cell-type annotations for each cluster.
panglao <- fread("PanglaoDB_markers_27_Mar_2020.tsv.gz", header = TRUE) %>%
  rename_with(~ gsub(" ", "_", .))
# QC for PanglaoDB
missing_summary <- panglao %>%
  summarise(across(everything(), ~ sum(is.na(.)))) %>%
  gather(column, missing_count)
marker_quality <- panglao %>%
  group_by(cell_type) %>%
  summarise(
    # Number of markers per cell type
    marker_count = n(),
    # Average specificity (combine human and mouse)
    avg_specificity = mean((specificity_human + specificity_mouse) / 2, na.rm = TRUE),
    # Average sensitivity
    avg_sensitivity = mean((sensitivity_human + sensitivity_mouse) / 2, na.rm = TRUE),
    # Count of canonical markers
    canonical_markers = sum(canonical_marker == 1, na.rm = TRUE)
  )
knitr::kable(marker_quality)
```

cell_type	marker_count	avg_specificity	avg_sensitivity	canonical_markers
Acinar cells	52	0.0109114	0.4190952	51
Adipocyte progenitor cells	23	0.0057907	0.0000000	8
Adipocytes	120	0.0193774	0.1451637	106
Adrenergic neurons	8	0.0031392	0.1250000	8
Airway epithelial cells	4	0.0212386	0.0000000	4
Airway goblet cells	30	0.0167041	0.1961207	10
Airway smooth muscle cells	4	0.0104603	0.0000000	0
Alpha cells	48	0.0185658	0.2910003	42
Alveolar macrophages	33	0.0232390	0.2291374	8
Anterior pituitary gland cells	36	0.0150746	0.1301020	12
Astrocytes	63	0.0238427	0.2229942	62
B cells	110	0.0258542	0.1903725	51
B cells memory	66	0.0181857	0.0000000	66
B cells naive	69	0.0273572	0.1362319	66
Basal cells	56	0.0382323	0.3749404	9
Basophils	82	0.0207592	0.1172996	82

cell_type	marker_count	avg_specificity	avg_sensitivity	canonical_markers
Bergmann glia	41	0.0320165	0.1707317	40
Beta cells	54	0.0148054	0.3124098	53
Cajal-Retzius cells	82	0.0140289	0.1399639	7
Cardiac stem and precursor cells	18	0.0087241	0.0000000	8
Cardiomyocytes	106	0.0194174	0.1233333	96
Cholangiocytes	43	0.0247460	0.2137209	43
Cholinergic neurons	7	0.0153732	0.0000000	7
Chondrocytes	52	0.0357844	0.1581699	46
Choroid plexus cells	22	0.0091985	0.1571146	9
Chromaffin cells	25	0.0134311	0.0000000	25
Ciliated cells	18	0.0166137	0.2222222	3
Clara cells	24	0.0078687	0.1928192	24
Crypt cells	16	0.0112396	0.0000000	16
Decidual cells	10	0.0288034	0.6000000	0
Delta cells	34	0.0110231	0.2408088	34
Dendritic cells	133	0.0390322	0.2002293	127
Distal tubule cells	49	0.0072426	0.1083777	14
Dopaminergic neurons	21	0.0092963	0.0000000	8
Ductal cells	43	0.0357928	0.1744186	42
Embryonic stem cells	85	0.0377222	0.1482684	64
Endothelial cells	195	0.0251803	0.2115172	192
Endothelial cells (aorta)	96	0.0106712	0.2653509	2
Endothelial cells (blood brain barrier)	17	0.0129859	0.0000000	0
Enteric glia cells	17	0.0240200	0.2864583	8
Enteric neurons	69	0.0098334	0.2057166	6
Enterochromaffin cells	12	0.0077238	0.0000000	12
Enterocytes	154	0.0127084	0.2227790	102
Enteroendocrine cells	48	0.0261081	0.1434742	48
Eosinophils	30	0.0116273	0.1206897	29
Ependymal cells	59	0.0113908	0.3151725	34
Epiblast cells	18	0.0454028	0.0000000	18
Epithelial cells	143	0.0195124	0.1127703	13
Epsilon cells	47	0.0179166	0.0000000	47
Erythroblasts	26	0.0502920	0.0000000	24
Erythroid-like and erythroid precursor cells	82	0.0313373	0.1948902	32
Fibroblasts	179	0.0316478	0.2135635	129
Follicular cells	14	0.0033972	0.2142857	8
Foveolar cells	24	0.0119663	0.3049242	7
GABAergic neurons	15	0.0212205	0.0833333	13
Gamma (PP) cells	29	0.0140324	0.1330049	29
Gamma delta T cells	66	0.0442583	0.1751042	40
Gastric chief cells	10	0.0052142	0.0000000	9
Germ cells	170	0.0069819	0.3321516	44
Glomus cells	28	0.0418535	0.0000000	28
Glutaminergic neurons	10	0.0242246	0.0000000	7
Glycinergic neurons	2	0.0003916	0.0000000	2
Goblet cells	37	0.0205149	0.3746346	16
Granulosa cells	18	0.0230478	0.2500000	6
Hemangioblasts	3	0.0033939	0.0000000	0
Hematopoietic stem cells	88	0.0417593	0.0939276	88
Hepatic stellate cells	46	0.0472446	0.2833333	21

cell_type	marker_count	avg_specificity	avg_sensitivity	canonical_markers
Hepatoblasts	17	0.0352874	0.0000000	9
Hepatocytes	154	0.0224800	0.2006668	135
His bundle cells	3	0.0087113	0.0000000	0
Immature neurons	38	0.0143608	0.0789474	37
Intercalated cells	25	0.0050308	0.1766667	25
Interneurons	221	0.0133377	0.0732935	49
Ionocytes	12	0.0064732	0.1666667	8
Juxtaglomerular cells	8	0.0094027	0.1904762	1
Keratinocytes	48	0.0203994	0.2493311	40
Kidney progenitor cells	7	0.0061012	0.5357143	0
Kupffer cells	49	0.0219675	0.1471354	42
Langerhans cells	30	0.0261280	0.1162879	17
Leydig cells	56	0.0180275	0.2119835	36
Loop of Henle cells	53	0.0158696	0.0384615	44
Luminal epithelial cells	48	0.0603353	0.3600427	48
Luteal cells	28	0.0224518	0.2916667	3
Macrophages	153	0.0202311	0.2169158	152
Mammary epithelial cells	46	0.0262343	0.2265396	31
Mast cells	162	0.0200734	0.1154231	144
Megakaryocytes	48	0.0358656	0.0000000	48
Melanocytes	42	0.0225771	0.1083333	32
Meningeal cells	18	0.0364349	0.1623932	11
Merkel cells	16	0.0532009	0.0000000	11
Mesangial cells	57	0.0494815	0.1798246	33
Mesothelial cells	58	0.0131628	0.2681287	4
Microfold cells	34	0.0878111	0.0000000	34
Microglia	80	0.0283793	0.1533071	65
Monocytes	102	0.0431124	0.2216950	70
Motor neurons	14	0.0038715	0.0000000	2
Myeloid-derived suppressor cells	17	0.0237091	0.0000000	2
Myoblasts	33	0.0147807	0.1255411	29
Myocytes	90	0.0109547	0.0802469	1
Myoepithelial cells	26	0.0223106	0.3108974	5
Myofibroblasts	9	0.0564886	0.2407407	4
Müller cells	54	0.0796667	0.1569017	45
NK cells	98	0.0182727	0.1976888	94
Natural killer T cells	24	0.0462988	0.0000000	24
Neural stem/precursor cells	58	0.0372501	0.0683270	28
Neuroblasts	50	0.0272722	0.0700000	50
Neuroendocrine cells	27	0.0182570	0.1095147	4
Neurons	211	0.0200332	0.1682849	81
Neutrophils	80	0.0170600	0.1641601	73
Noradrenergic neurons	8	0.0023201	0.1250000	8
Nuocytes	12	0.0207915	0.1944444	8
Olfactory epithelial cells	136	0.0141732	0.1891791	1
Oligodendrocyte progenitor cells	28	0.0271361	0.3618848	28
Oligodendrocytes	88	0.0087524	0.2534838	88
Osteoblasts	64	0.0216746	0.1108530	62
Osteoclast precursor cells	8	0.0068763	0.0000000	3
Osteoclasts	49	0.0237483	0.1173469	49
Osteocytes	5	0.1731954	0.0000000	5
Oxyphil cells	2	0.0007389	0.0000000	2

cell_type	marker_count	avg_specificity	avg_sensitivity	canonical_markers
Pancreatic progenitor cells	15	0.0539997	0.0000000	13
Pancreatic stellate cells	29	0.0668016	0.5427019	19
Paneth cells	54	0.0113609	0.1416274	49
Parathyroid chief cells	8	0.0107731	0.1666666	8
Parietal cells	14	0.0146750	0.0000000	1
Peri-islet Schwann cells	16	0.0049590	0.5343750	1
Pericytes	64	0.0232344	0.2403534	39
Peritubular myoid cells	26	0.0404972	0.1883019	7
Photoreceptor cells	55	0.0094775	0.1514499	3
Pinealocytes	24	0.0102331	0.1458333	24
Plasma cells	86	0.0298742	0.1723191	86
Plasmacytoid dendritic cells	58	0.0287719	0.2727671	53
Platelets	131	0.0446603	0.2417303	131
Pluripotent stem cells	21	0.0387148	0.0000000	19
Podocytes	95	0.0266004	0.2535353	93
Principal cells	44	0.0211838	0.2109634	19
Proximal tubule cells	83	0.0131336	0.2150000	81
Pulmonary alveolar type I cells	36	0.0184713	0.2357026	36
Pulmonary alveolar type II cells	47	0.0137582	0.3175408	46
Pulmonary vascular smooth muscle cells	2	0.0210085	0.0000000	0
Purkinje fiber cells	5	0.0120309	0.0000000	0
Purkinje neurons	60	0.0148246	0.0813008	7
Pyramidal cells	35	0.0324010	0.0857143	27
Radial glia cells	14	0.0132390	0.1785714	14
Red pulp macrophages	12	0.0104112	0.1623563	12
Reticulocytes	8	0.0200048	0.0000000	8
Retinal ganglion cells	70	0.0081243	0.3580655	5
Retinal progenitor cells	13	0.0191938	0.0940171	7
Salivary mucous cells	14	0.0124348	0.3685186	0
Satellite cells	44	0.0261598	0.0730897	37
Satellite glial cells	24	0.0320804	0.0000000	24
Schwann cells	48	0.0336488	0.1219125	39
Sebocytes	25	0.0271756	0.0000000	25
Serotonergic neurons	8	0.0030015	0.3125000	7
Sertoli cells	72	0.0186404	0.1210908	70
Smooth muscle cells	82	0.0140240	0.1888048	80
Spermatocytes	22	0.0195569	0.0759943	0
Spermatozoa	9	0.0191995	0.0000000	0
Stromal cells	35	0.0189855	0.0371429	6
T cells	107	0.0329686	0.1805976	51
T cells naive	4	0.0382405	0.3750000	0
T cytotoxic cells	8	0.0330086	0.2500000	2
T follicular helper cells	13	0.0174711	0.0000000	3
T helper cells	62	0.0127760	0.0000000	26
T memory cells	63	0.0321186	0.3250943	3
T regulatory cells	21	0.0162962	0.0595238	13
Tanycytes	27	0.0436107	0.1119929	27
Taste receptor cells	20	0.0066867	0.0000000	19
Thymocytes	23	0.0252609	0.0000000	23
Transient cells	23	0.0163187	0.0000000	2
Trichocytes	6	0.0061080	0.0000000	6
Trigeminal neurons	46	0.0102348	0.2435771	3

cell_type	marker_count	avg_specificity	avg_sensitivity	canonical_markers
Trophoblast cells	31	0.0202157	0.0000000	5
Trophoblast progenitor cells	3	0.0289949	0.0000000	3
Trophoblast stem cells	1	0.0032895	0.0000000	1
Tuft cells	38	0.0403882	0.1794872	38
Undefined placental cells	22	0.0181862	0.3083333	0
Urothelial cells	11	0.0020182	0.0568182	11
Vascular smooth muscle cells	6	0.0152298	0.2291667	0

```
filtered_markers <- panglao %>%
  filter(
    species %in% c("Mm Hs", "Hs"),
    (specificity_human > 0.01 | specificity_mouse > 0.01),
    ubiquitousness_index < 0.05,
    (sensitivity_human > 0.1 | sensitivity_mouse > 0.1),
    !is.na(canonical_marker) # We don't restrict organs.
  )
# Function to select a gene column and get unique rows.
# Selection by "Quality Score".
select_unique_genes <- function(markers, gene_col_name) {
  markers %>%
    mutate(
      quality_score = (specificity_human + specificity_mouse +
        sensitivity_human + sensitivity_mouse) / 4 *
        if_else(canonical_marker == 1, 1.5, 1)
    ) %>%
    group_by(across(all_of(gene_col_name))) %>%
    slice_max(order_by = quality_score, n = 1, with_ties = FALSE) %>%
    ungroup() %>%
    as.data.table()
}
unique_markers <- select_unique_genes(filtered_markers, "official_gene_symbol")
```

We will also consider matching by nicknames if we cannot find official gene names.

```
# Expand Nicknames to be on seperate rows.
expand_nicknames <- function(dt) {
  result <- copy(dt)
  result <- result[, .(
    nickname = unlist(strsplit(nicknames, "\\|"))
  ), by = .(
    species, official_gene_symbol, cell_type, ubiquitousness_index,
    product_description, gene_type, canonical_marker, germ_layer,
    organ, sensitivity_human, sensitivity_mouse, specificity_human,
    specificity_mouse, quality_score
  )]
  return(result)
}
unique_nickname_marker <- expand_nicknames(unique_markers)
unique_nickname_marker <- select_unique_genes(unique_nickname_marker, "nickname") %>%
  filter(!is.na(nickname))
```

The Matching:

```
all_markers <- fread("scrna_all_markers.txt", header = TRUE)
annotated_markers <- all_markers %>%
  # First try to match with official gene symbols
  left_join(
    unique_markers %>%
      select(
        official_gene_symbol, cell_type,
        specificity_human, specificity_mouse,
```

```
        sensitivity_human, sensitivity_mouse
      ),
    by = c("gene" = "official_gene_symbol")
  ) %>%
  # For rows where we didn't find a match (where cell_type is NA),
  # try to match with nicknames
  left_join(
    unique_nickname_marker %>%
      select(
        nickname, cell_type,
        specificity_human, specificity_mouse,
        sensitivity_human, sensitivity_mouse
      ),
    by = c("gene" = "nickname")
  ) %>%
  mutate(
    cell_type = coalesce(cell_type.x, cell_type.y),
    specificity_human = coalesce(specificity_human.x, specificity_human.y),
    specificity_mouse = coalesce(specificity_mouse.x, specificity_mouse.y),
    sensitivity_human = coalesce(sensitivity_human.x, sensitivity_human.y),
    sensitivity_mouse = coalesce(sensitivity_mouse.x, sensitivity_mouse.y)
  ) %>%
  select(-ends_with(".x"), -ends_with(".y"))
```

Check Matching Summary:

```
na_count <- sum(is.na(annotated_markers$cell_type))
total_rows <- nrow(annotated_markers)
print(paste0("Number of NAs in cell_type: ", na_count))
```

```
## [1] "Number of NAs in cell_type: 21158"
```

```
print(paste0("Total number of rows: ", total_rows))
```

```
## [1] "Total number of rows: 23699"
```

```
print(paste0("Percentage of NAs: ", round(na_count / total_rows * 100, 2), "%"))
```

```
## [1] "Percentage of NAs: 89.28%"
```

```
write.table(annotated_markers, "scrna_annotated_markers.txt",
  quote = FALSE,
  row.names = FALSE, col.names = TRUE, sep = "\t"
)
```