scripts

January 5, 2024

```python
#!/usr/bin/env python3

# scripts.py

# Notebook of scripts for analysis

# Vivian Leung
# Created:      05 Jan 2024
# Last updated: 05 Jan 2024
# Last used:    05 Jan 2024

# Changelog:

# %%
# IMPORTS
import os

import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```python
# USER PARAMS

# I/O
PROJECT_DIR = "."

DATA_DIR = f"{PROJECT_DIR}/data"

# Input
COUNTS_TSV = f"{DATA_DIR}/processing_counts.tsv"

# Output
OUT_DIR = f"{PROJECT_DIR}/analysis"

# Params
```

```python
# FUNCTIONS
```

```
[ ]: # SCRIPT

     counts = pd.read_csv(COUNTS_TSV, sep="\t")

     # cast as ordered for easy sorting in figs
     sample_name_dtype = pd.CategoricalDtype(counts.sort_values("order").
       ↪sample_name, True)
     counts["sample_name"] = counts["sample_name"].astype(sample_name_dtype)

     # organize
     counts.set_index("sample_name", inplace=True)

     counts["read1_filter_excluded"] = counts.read1_raw - counts.read1_filtered
     counts["read2_filter_excluded"] = counts.read2_raw - counts.read2_filtered
```

```
[ ]: rename_cols = {
         "read1_raw": "raw",
         "read1_filtered": "filtered",
         "read1_filter_excluded": "removed by filter",
         "seqs_classified": "classified",
         "seqs_unclassified": "unclassified",
     }
     steps = counts[rename_cols.keys()].rename(rename_cols, axis=1).rename_axis(
         "read pairs", axis=1)

     steps.style.format("{:,.0f}")

     ax_steps_counts = steps.drop(["raw", "filtered"], axis=1).plot(
         kind="bar",
         stacked=True,
         xlabel="Sample",
         ylabel="Number of read pairs ($10^7$)",
         title="Read data processing",
     )
     ax_steps_counts.legend_ = None
     ax_steps_counts.legend(loc="upper left", bbox_to_anchor=(1, 1))
```
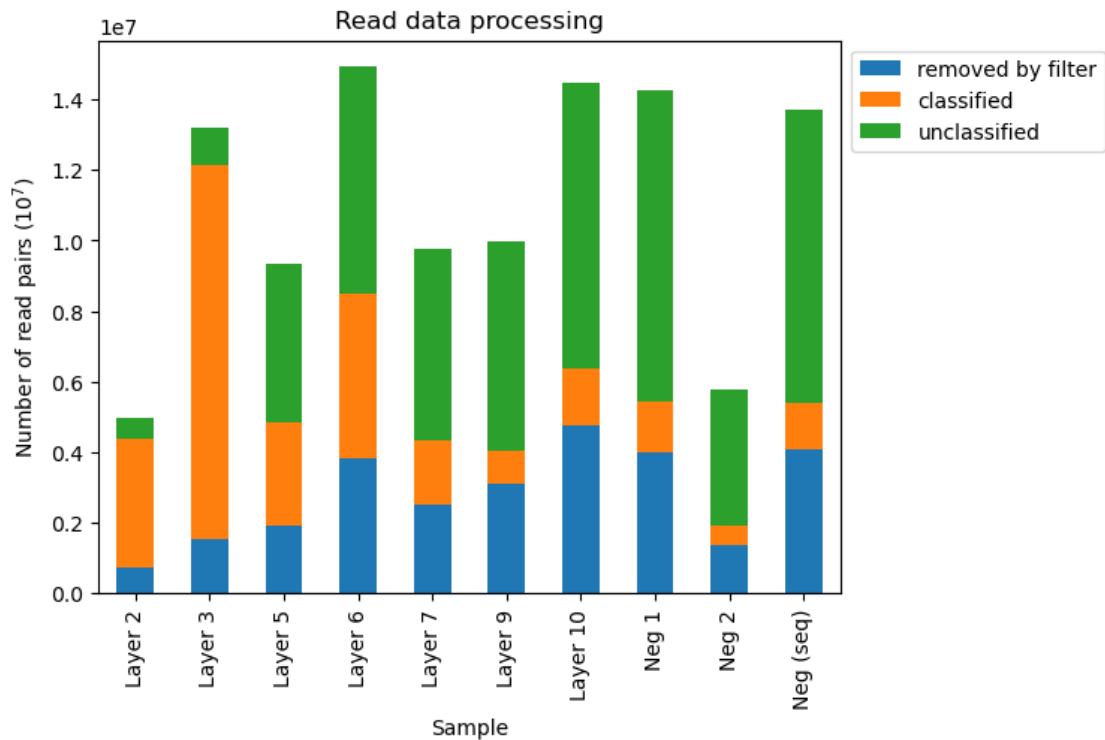
[ ]: <matplotlib.legend.Legend at 0x1aa536cc0>
```

```
steps_pcts = steps.apply(lambda ser: ser / ser["raw"], axis=1)

steps_pcts.style.format("{:.2f}")


ax_steps_pcts = steps_pcts.drop(["raw", "filtered"], axis=1).plot(
    kind="bar",
    stacked=True,
    xlabel="Sample",
    ylabel="Proportion of read pairs",
    title="Read data processing",
)

# move legend
ax_steps_pcts.legend_ = None
ax_steps_pcts.legend(loc="upper left", bbox_to_anchor=(1, 1))
```

[ ]: <matplotlib.legend.Legend at 0x1aa52edb0>

Read data processing