Contents

This is a script for annotating an array given it's probe-gene table and a gff.

1 Code

1.1 Create rosetta

Rosetta is a dictionary with information of gene names and annotation. It is created merging info from:

- The gff (from plasmoDB)
- A file containing gene aliases (from PlasmoDB): $aliases_{file}$.
- A file containing "names" of the genes: $gene_{namesfile}$.

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

rosetta = {}
with open("/home/lucas/ISGlobal/Gen_Referencies/Gene_references_rosetta.txt", "r+") as
    for line in file1:
        rosetta[str(line.split("\t")[0].strip())] = {
            "old_refs": line.split("\t")[1:]}
    for key, value in rosetta.items():
        value["old_refs"][-1] = value["old_refs"][-1].strip()

with open("/home/lucas/ISGlobal/Gen_Referencies/PlasmoDB-41_Pfalciparum3D7.gff", "r+")
    for line in file2:
        if line.startswith("#"):
            pass
        elif line.split()[2] == "gene":
            line_split = line.strip().split("\t")[8].split(";")
            if line_split[0].replace("ID=", "") in rosetta.keys():
                rosetta[line_split[0].replace("ID=", "")]["annot"] = line_split[1].rep
                    "description=", "")
        else:
            pass

with open("/home/lucas/ISGlobal/Gen_Referencies/gene_names.txt", "r+") as file3:
    header = True
    for line in file3:
        if header:
            pass
            header = False
        else:
            if line.strip().split()[0] in rosetta.keys():
                if line.strip().split("\t")[4] == "N/A":
                    rosetta[line.strip().split("\t")[0]]["name"] = "NA"
                else:
                    rosetta[line.strip().split("\t")[0]
                            ]["name"] = line.strip().split("\t")[4]
```

1.2 Load array to gene mapping and status

Load the description of the array:

- Probenames
- Target gene for each probe
- Whether it should be kept (we remove probes that map to multiple genes).
- We add annotation for the new probes and GDV1

```
import collections as col
import re

array_dict = col.defaultdict(dict)
with open("/media/lucas/Disc4T/Projects/Microarrays_R_analysis/array_decription.csv") a
    for line in infile:
        line_list = line.strip().split()
        probe = line_list[1]
        gene = line_list[3]
        status = line_list[4]
        array_dict[probe] = {"gene":gene, "status":status}

for k,v in array_dict.items():
    if k.startswith("PF3D7"):
        v["gene"] = re.sub(r'_n\d.*', "", k)

for k,v in array_dict.items():
    if k.startswith("gdv1"):
        v["gene"] = "PF3D7_0935400"
```

1.3 Load array info

```
with open("/media/lucas/Disc4T/Projects/Oriol/Microarrays/Raw_Data/US10283823_258576310

    skip = 10
    i = 1

    for line in infile:
        if i > skip:

            probe = line.split()[6]
            gene = array_dict[probe]["gene"]
            status = array_dict[probe]["status"]

            try:
                name = rosetta[gene]["name"]
                anot = rosetta[gene]["annot"]
                print("\t".join([probe, gene, status, name, anot]))

            except:
                print("\t".join([probe, gene, status, gene, gene]))


        else:
            i += 1
```