Creation of active gene-lists

Lucas Michel Todó

31/03/2020

Contents

1 Active/Inactive Gene Lists

Our aim is to create a unified table that assigns to each gene in the P.falicparum gnome a expresion state. We will define 4 possible expresion states:

- Active

  - Regular
  - Variant Active
  - Variant Repressed

- Inactive

1.1 Load Packages and functions

1.2 Microarray Data: Red Signal

We willl load the red signal and transform it into percentiles. For each gene we pick the "Aver.2Higher" column from the original microarrays data table. This column corresponds to the average between the two highest red signals among available timepoints.

Red Signal DataFrame

```
               Gene_id      Red_12B      Red_10G      Red_3D7B Percent_12B
1           mal_mito_3 22579.33333 36436.73333 30636.82500  96.0335622
2 MAL13P1.415_oldname    770.82083    702.22292    640.11667  21.3196034
3  MAL13P1.65_oldname    111.33333     87.05833     91.05833   6.2166285
4  MAL7P1.142_oldname   5924.44167   5194.40000   5114.63333  75.4767353
5  MAL8P1.310_oldname     37.21250     35.37917     33.24167   0.8581236
6   MAL8P1.90_oldname     80.55417     46.18333     54.64167   4.1952708
  Percent_10G Percent_3D7B MaxRedPercentDif
1   98.474447   97.8832952        2.4408848
2   20.861937   18.4591915        2.8604119
3    5.053394    4.5194508        1.6971777
4   72.444699   71.7200610        3.7566743
5    1.115561    0.5911518        0.5244088
6    2.002288    2.2501907        2.1929825
```

1.3 Microarray Data: Areas

We will load the areas data to calculate FC among strains. For each gene, we select the time interval (right, left, mid or sides) for which we find the

maximum difference among strains (between highest and lowest). We will also add a column to check if this time interval corresponds to the interval of maxium expression for each strain.

Areas DataFrame

```
               Gene_id      l_12B     r_12B     m_12B    s_12B     l_10G    r_10G
1            mal_mito_3 30.592496 61.080128 49.676556 41.99607 25.372470 62.38873
2 MAL13P1.415_oldname  5.423269  8.488971  1.289779 12.62246  6.117132 10.59524
3  MAL13P1.65_oldname 18.322430        NA 17.593468       NA 14.071128       NA
4  MAL7P1.142_oldname  9.389247 12.807814 10.340803 11.85626 13.661078 14.52676
5  MAL8P1.310_oldname        NA        NA        NA       NA        NA       NA
6   MAL8P1.90_oldname        NA        NA        NA       NA        NA       NA
      m_10G     s_10G     l_3D7B    r_3D7B    m_3D7B    s_3D7B   MaxLeft   MinLeft
1 49.805504 37.95570 25.484634 62.83441 50.462696 37.856349 30.592496 25.372470
2  3.676218 13.03616  1.789873 10.51234  3.691753  8.610459  6.117132  1.789873
3        NA       NA 19.333324       NA        NA        NA 19.333324 14.071128
4 13.401610 14.78623  7.099032 13.34518 12.041177  8.403034 13.661078  7.099032
5        NA       NA        NA       NA        NA        NA        NA       NA
6        NA       NA        NA       NA        NA        NA        NA       NA
  MaxRight  MinRight    MaxMid    MinMid MaxSides  MinSides  DifLeft DifRight
1 62.83441 61.080128 50.462696 49.676556 41.99607 37.856349 5.220025 1.754283
2 10.59524  8.488971  3.691753  1.289779 13.03616  8.610459 4.327259 2.106274
3       NA        NA        NA        NA       NA        NA 5.262196       NA
4 14.52676 12.807814 13.401610 10.340803 14.78623  8.403034 6.562046 1.718950
5       NA        NA        NA        NA       NA        NA       NA       NA
6       NA        NA        NA        NA       NA        NA       NA       NA
      DifMid DifSides Interval      Dif
1 0.7861401 4.139719  DifLeft 5.220025
2 2.4019733 4.425700 DifSides 4.425700
3        NA       NA  DifLeft 5.262196
4 3.0608067 6.383199  DifLeft 6.562046
5        NA       NA  No Data       NA
6        NA       NA  No Data       NA
```

1.4   Load RNA-Seq Data

We will use publicly available data from PlasmoDB to create a reference expresion percentile for each gene. All data-sets are from RNA-Seq studies in the 3D7 strain. We are using 4 different data-sets:

- Otto et.al.

- Hoeijmakers et.al.

- Toenhake et.al.

- Bartfai et.al.

```
       Gene_id MaxPercOtto MaxPercHoej MaxPercToen MaxPercBart MeanPercent
1 PF3D7_0100100        57.2        54.3        33.9        31.7      44.275
2 PF3D7_0100200        29.4        50.5        26.6        36.0      35.625
3 PF3D7_0100300        34.2         8.7         7.7         7.4      14.500
4 PF3D7_0100400        50.3        18.3        11.3        37.4      29.325
5 PF3D7_0100500        49.7        11.4        14.0        32.5      26.900
6 PF3D7_0100600        18.5         7.8         2.3        12.1      10.175
  StdDevPercent
1     11.546942
2      9.241313
3     11.383980
4     15.424068
5     15.474657
6      5.930167
```

We plot the standard deviation of the percentile values among different studies and we can see that for the vast majority of genes it doesn't go above 10.

SD of Percent value (among different studies)
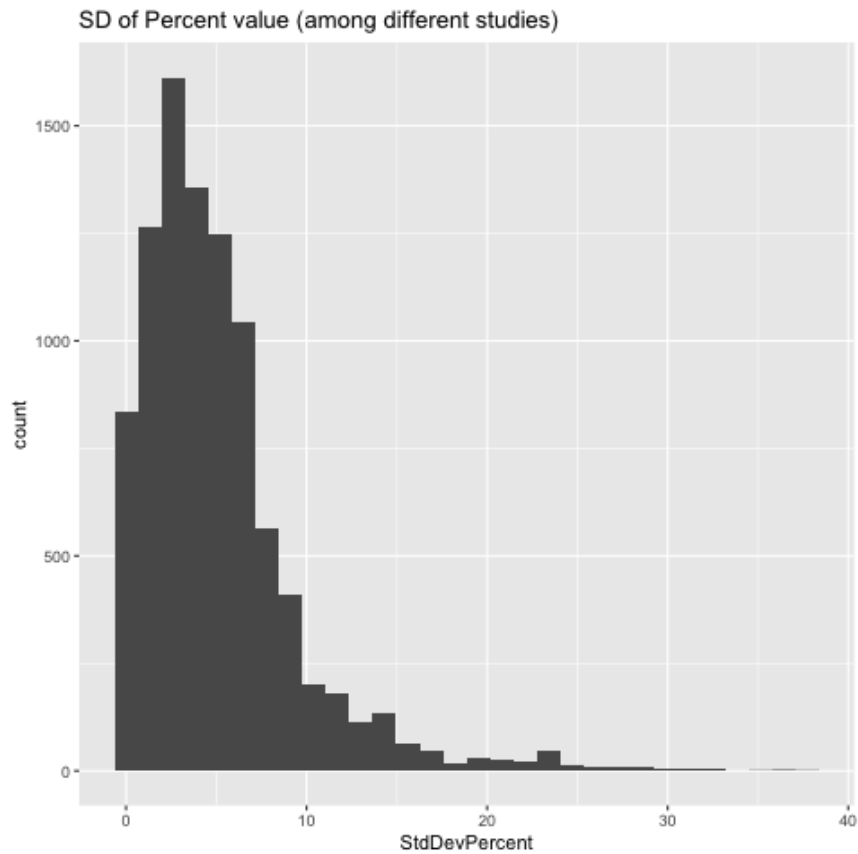
1.5   Create Join DF

`org_babel_R_eoe`

1.6   Create Lists according to thresholds

Now that we have all the data loaded in, we can star to set labels for each gene.

We will jhave the following categories:

- Active : Mean percentile (from rna-seq experiments) $>= 25\%$

    - Regular : Non-variant according to our variant genes list.
    - Variant Active: Variant according to our variant genes list.
    - Variant Repressed: Variant and red signal percentile differnce $>$ 30% and area diffrence > 1 (~2FC)

- Inactive: Mean percentile (from rna-seq experiments) $< 25\%$