summary_file

My Linh Thibodeau

2017-11-11

Library and reading tables

```
suppressMessages(suppressWarnings(library(tidyverse)))
knitr::opts_chunk$set(fig.width=12, fig.height=9)
library(knitr)
library(kableExtra)
```

```
## Warning: package 'kableExtra' was built under R version 3.4.2
```

```
options(knitr.table.format = "html")
mut_sig <- read.table("mut_sig.tsv", header = TRUE, sep = "\t")
ref_mut_sig_ordered_by_sig3 <- readRDS("somatic_mutations_formated_files/ref_mut_sig_ordered_by_sig3.rds
mut_sig_gather <- read.table("mut_sig_gather.tsv", header = TRUE, sep="\t")
ALL_mut_gather <- read.table("somatic_mutations_formated_files/ALL_mut_gather.tsv", header = TRUE, sep =
aml_mut_gather <- read.table("somatic_mutations_formated_files/aml_mut_gather.tsv", header = TRUE, sep =
breast_mut_gather <- read.table("somatic_mutations_formated_files/breast_mut_gather.tsv", header = TRUE
medullo_mut_gather <- read.table("somatic_mutations_formated_files/medulloblastoma_mut_gather.tsv", hea
pancreas_mut_gather <- read.table("somatic_mutations_formated_files/pancreas_mut_gather.tsv", header = T
all_cancer_types_mut <- read.table("somatic_mutations_formated_files/all_cancer_types_mut.tsv", header =
all_cancer_types_mut_with_ref_sig <- read.table("somatic_mutations_formated_files/all_cancer_types_mut_w
all_cancer_mutations_per_snv_sig_score <- read.table("somatic_mutations_formated_files/all_cancer_mutati
all_cancer_types_mut_proportion_signatures <-  read.table("somatic_mutations_formated_files/all_cancer_t
all_cancer_types_stats <- read.table("statistics/all_cancer_types_stats.tsv", header = TRUE, sep = "\t")
```

DATA INFORMATION

For this homework, I will be using open access complete set of cancer somatic mutations from:

Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. 500, 415–421 (2013).

Please note that I deliberately broke down each step of this homework in very small chunks/steps/scripts to make this material more "generalizable", which will allow me to use them for my research work as well.

This files will contain a narrative of homework 7 and some helpful notes regarding my process.

Download the data

I started by using curl shell script in my Makefile to download the complete set of cancer somatic mutations (see Makefile for details):

```
mut_sig_raw.txt:
    curl -o mut_sig_raw.txt ftp://ftp.sanger.ac.uk/pub/cancer/AlexandrovEtAl/signatures.txt
```

I am also using my `mut_sig_clean.R` script to format the data. I performed that step to remove empty spaces from the column names because this would create problems with variable manipulation later.

Perform exploratory analyses

Tasks:

- Bring the data in as data frame.
- Save a couple descriptive plots to file with highly informative names.
- Reorder the continents based on life expectancy. You decide the details. Sort the actual data in a deliberate fashion. You decide the details, but this should at least implement your new continent ordering.
- Write the Gapminder data to file(s), for immediate and future reuse.

Please note that I did not complete the 5 tasks above in the order suggested because it made more sense to proceed in the order outlined below with my dataset !

The original format of the reference mutation signature is:

```
mut_sig %>% arrange(Somatic.Mutation.Type) %>% head(5) %>% kable()
```

Substitution.Type

Trinucleotide

Somatic.Mutation.Type

Signature.1A

Signature.1B

Signature.2

Signature.3

Signature.4

Signature.5

Signature.6

Signature.7

Signature.8

Signature.9

Signature.10

Signature.11

Signature.12

Signature.13

Signature.14

Signature.15

Signature.16

Signature.17

Signature.18

Signature.19

Signature.20

Signature.21

Signature.R1

Signature.R2

Signature.R3

Signature.U1

Signature.U2

C>A

ACA

A[C>A]A

0.0112

0.0104

0.0105

0.0240

0.0365

0.0149

0.0017

4e-04

0.0368

0.0120

0.0007

2e-04

0.0077

0.0007

0.0001

0.0013

0.0161

0.0018

0.0500

0.0107

0.0013

1e-04

0.0210

0.0137

0.0044

0.0105

0.0221
C>A
ACC
A[C>A]C
0.0092
0.0093
0.0061
0.0197
0.0309
0.0089
0.0028
5e-04
0.0287
0.0067
0.0010
1e-03
0.0047
0.0001
0.0042
0.0040
0.0097
0.0003
0.0076
0.0074
0.0024
7e-04
0.0065
0.0046
0.0047
0.0005
0.0123
C>A
ACG
A[C>A]G
0.0015
0.0016

0.0013

0.0019

0.0183

0.0022

0.0005

0e+00

0.0017

0.0005

0.0003

0e+00

0.0017

0.0001

0.0005

0.0000

0.0022

0.0000

0.0017

0.0005

0.0000

0e+00

0.0000

0.0048

0.0003

0.0000

0.0028

C>A

ACT

A[C>A]T

0.0063

0.0067

0.0037

0.0172

0.0243

0.0092

0.0019

4e-04

0.0300

0.0068

0.0092

2e-04

0.0046

0.0002

0.0296

0.0057

0.0088

0.0032

0.0181

0.0074

0.0029

6e-04

0.0058

0.0081

0.0034

0.0112

0.0118

C>G

ACA

A[C>G]A

0.0018

0.0051

0.0048

0.0216

0.0097

0.0117

0.0013

0e+00

0.0085

0.0048

0.0005

7e-04

0.0031

0.0018

0.0001

0.0011

0.0048

0.0016

0.0014

0.0058

0.0005

5e-04

0.0038

0.0024

0.0012

0.0044

0.0108

(1) I am using my `mut_sig_tables.Rmd` script bring the mutational signatures in as a data frame. I am using the gather function of tidyverse to create a new dataframe table which will facilitate making plots later. I saved the output of this new data frame to mut_sig_gather.tsv.

```
mut_sig_gather %>% arrange(Somatic.Mutation.Type) %>% head(5) %>% kable()
```

Substitution.Type

Trinucleotide

Somatic.Mutation.Type

Signature

Score

C>A

ACA

A[C>A]A

Signature.1A

0.0112

C>A

ACA

A[C>A]A

Signature.1B

0.0104

C>A

ACA

A[C>A]A

Signature.2

0.0105

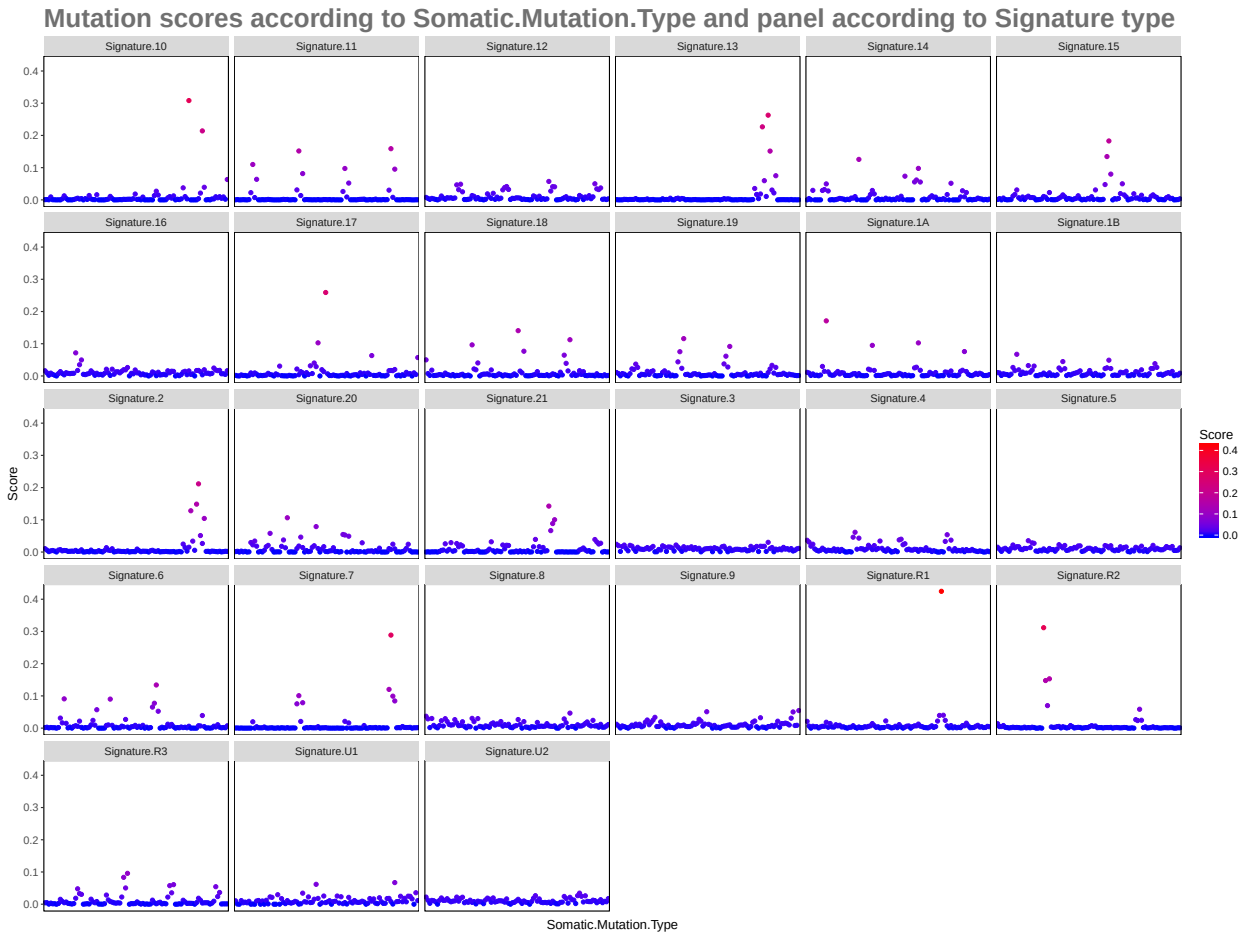


Figure 1: mutation_scores_for_all_snv_facet_signature

C>A

ACA

A[C>A]A

Signature.3

0.0240

C>A

ACA

A[C>A]A

Signature.4

0.0365

Note. This allows us to know how much each Somatic.Mutation.Type contributes to each signature.

(2) Here are some initial plots

(3) In this section, I am using the mut_sig_plot.Rmd script to create some plots, which I save in the subfolder called plots. I am reordering the Somatic.Mutation.Type (e.g. A[C>A]A, A[C>A]C, etc.)

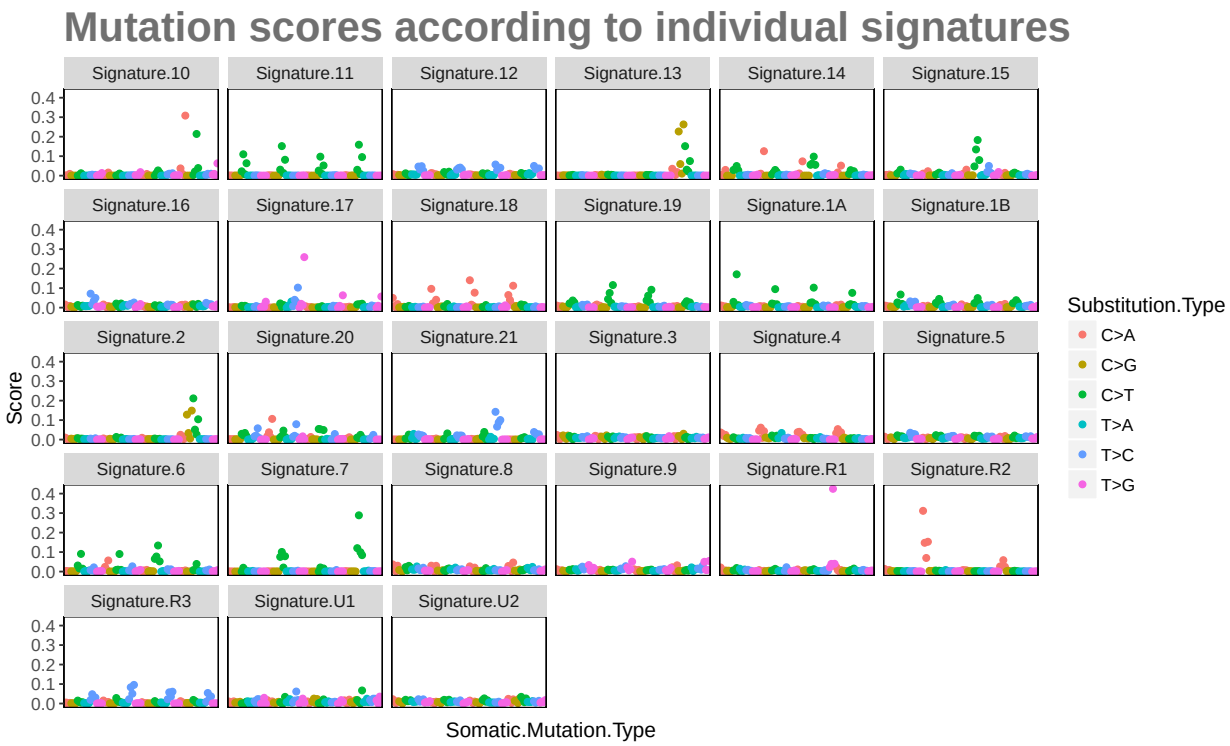


Figure 2: mutation_scores_for_all_snv_type_facet_signature_colored_snv

according to their Signature.3 Score.

```
ref_mut_sig_ordered_by_sig3 %>% head(5) %>% kable()
```

Substitution.Type

Trinucleotide

Somatic.Mutation.Type

Signature.1A

Signature.1B

Signature.2

Signature.3

Signature.4

Signature.5

Signature.6

Signature.7

Signature.8

Signature.9

Signature.10

Signature.11

Figure 3: cancer_type_sig_compare_plots

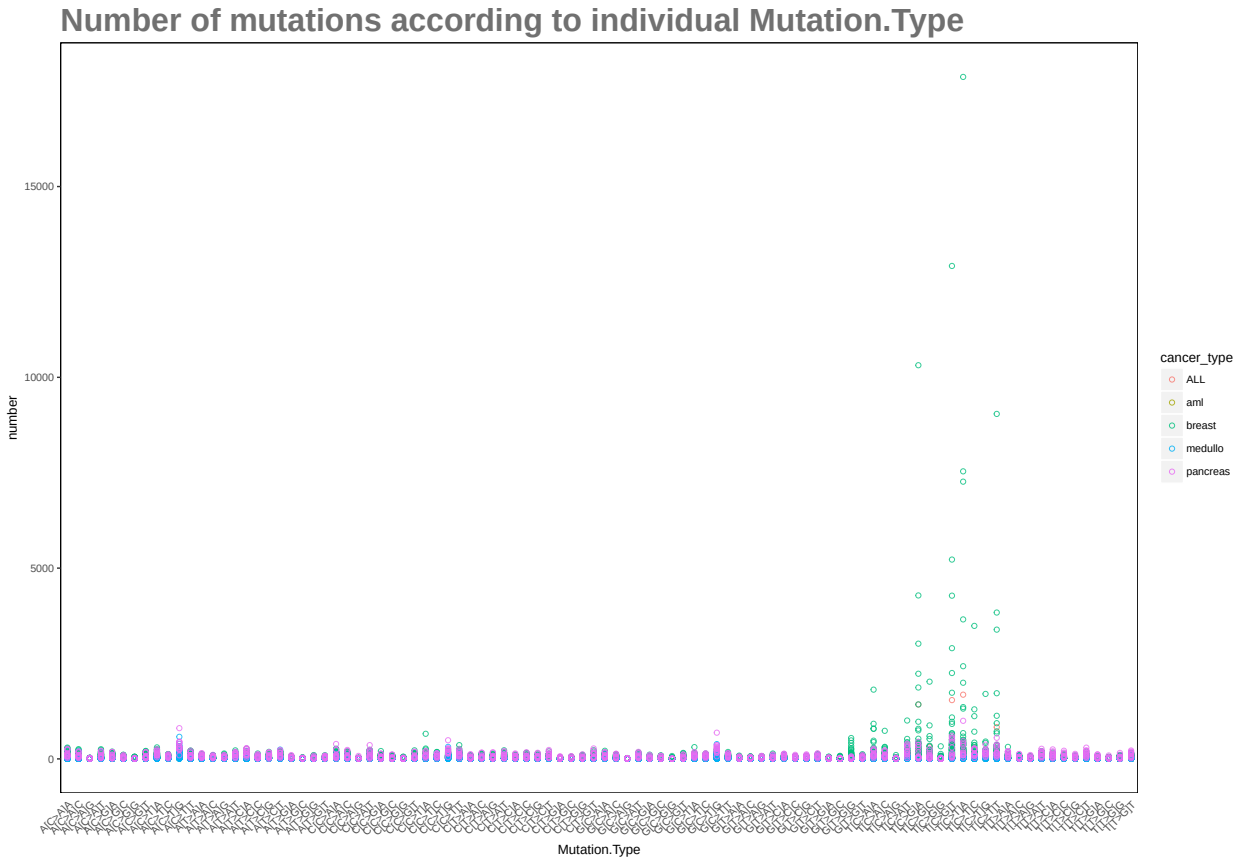


Figure 4: all_cancer_types_mut_geompoint

Signature.12

Signature.13

Signature.14

Signature.15

Signature.16

Signature.17

Signature.18

Signature.19

Signature.20

Signature.21

Signature.R1

Signature.R2

Signature.R3

Signature.U1

Signature.U2

C>A

ACA

A[C>A]A

0.0112

0.0104

0.0105

0.0240

0.0365

0.0149

0.0017

0.0004

0.0368

0.0120

0.0007

2e-04

0.0077

0.0007

0.0001

0.0013

0.0161

0.0018

0.0500

0.0107

0.0013

1e-04

0.0210

0.0137

0.0044

0.0105

0.0221

C>A

ACC

A[C>A]C

0.0092

0.0093

0.0061

0.0197

0.0309

0.0089

0.0028

0.0005

0.0287

0.0067

0.0010

1e-03

0.0047

0.0001

0.0042

0.0040

0.0097

0.0003

0.0076

0.0074

0.0024

7e-04

0.0065

0.0046

0.0047

0.0005

0.0123

C>A

ACG

A[C>A]G

0.0015

0.0016

0.0013

0.0019

0.0183

0.0022

0.0005

0.0000

0.0017

0.0005

0.0003

0e+00

0.0017

0.0001

0.0005

0.0000

0.0022

0.0000

0.0017

0.0005

0.0000

0e+00

0.0000

0.0048

0.0003

0.0000

0.0028

C>A

ACT

A[C>A]T

0.0063
0.0067
0.0037
0.0172
0.0243
0.0092
0.0019
0.0004
0.0300
0.0068
0.0092
2e-04
0.0046
0.0002
0.0296
0.0057
0.0088
0.0032
0.0181
0.0074
0.0029
6e-04
0.0058
0.0081
0.0034
0.0112
0.0118
C>A
CCA
C[C>A]A
0.0067
0.0090
0.0061
0.0194
0.0461
0.0097

0.0101

0.0012

0.0303

0.0098

0.0031

7e-04

0.0135

0.0035

0.0056

0.0106

0.0159

0.0010

0.0965

0.0112

0.0178

2e-03

0.0076

0.3117

0.0156

0.0173

0.0057

Note. We are still with the original format of the reference signatures here. I have decided to present only one example of ordering the data according to a factor, for more examples, you may refer to previous work from homework 5.

(4) Writing the mutational signatures input/output to files is performed (embedded) in the scripts mentioned above. I have read and written a lot of files in each script. Please refer to the section "Automate the pipeline" below.

Perform statistical analyses

Tasks:

- Import the data created in the first script.
- Make sure your new continent order is still in force. You decide the details.
- Fit a linear regression of life expectancy on year within each country. Write the estimated intercepts, slopes, and residual error variance (or sd) to file. The R package broom may be useful here.
- Find the 3 or 4 "worst" and "best" countries for each continent. You decide the details.

As you know, I am using genomic data, so I have to adapt the tasks to my data, so here is what I did:

(1) I am importing and cleaning the data with my script read_clean_genome_text_files.R:

- reference mutation signatures (mut_sig.txt)
- somatic mutations for 5 types of cancer: ALL, AML, breast, medulloblastoma, pancreas

I have performed the same "gathering" steps as described in the previous section for each cancer type.

```
ALL_mut_gather  %>% arrange(Mutation.Type) %>% head(5) %>% kable()
```

Mutation.Type

case_id

number

A[C>A]A

PD4020a

33

A[C>A]C

PD4020a

15

A[C>A]G

PD4020a

1

A[C>A]T

PD4020a

24

A[C>G]A

PD4020a

31

```
aml_mut_gather  %>% arrange(Mutation.Type) %>% head(5) %>% kable()
```

Mutation.Type

case_id

number

A[C>A]A

X400220

8

A[C>A]A

X426980

11

A[C>A]A

X452198

1

A[C>A]A

X573988

3

A[C>A]A

X758168

11

```
breast_mut_gather  %>% arrange(Mutation.Type) %>% head(5) %>% kable()
```

Mutation.Type

case_id

number

A[C>A]A

PD3851a

29

A[C>A]A

PD3890a

99

A[C>A]A

PD3904a

114

A[C>A]A

PD3905a

88

A[C>A]A

PD3945a

235

```
medullo_mut_gather  %>% arrange(Mutation.Type) %>% head(5) %>% kable()
```

Mutation.Type

case_id

number

A[C>A]A

LFS_MB1

45

A[C>A]A

LFS_MB2

23

A[C>A]A

LFS_MB4

24

A[C>A]A

MB1

6

A[C>A]A

MB101

82

```
pancreas_mut_gather  %>% arrange(Mutation.Type) %>% head(5) %>% kable()
```

Mutation.Type

case_id

number

A[C>A]A

APGI_1839

77

A[C>A]A

APGI_1840

156

A[C>A]A

APGI_1956

59

A[C>A]A

APGI_1992

117

A[C>A]A

APGI_2000

122

(2) The Somatic.Mutation.Type of the mut_sig.tsv (reference data) are still ordered according to Signature.3 values, as illustrated in Signature3_compare_plots.pdf

(3) Here, we are not looking at signatures of the reference somatic mutation (mut_sig.tsv) but we are looking at the Mutation.Type of 5 types of cancer: ALL, AML, breast, medulloblastoma, pancreas.

Summary statistics

We will be looking at the mean, median, standard deviation and count of each Mutation.Type in each dataset (we need the "gathered" versions of files for that step).

```
all_cancer_types_stats %>% head(10) %>% kable()
```

Mutation.Type

mean_aml

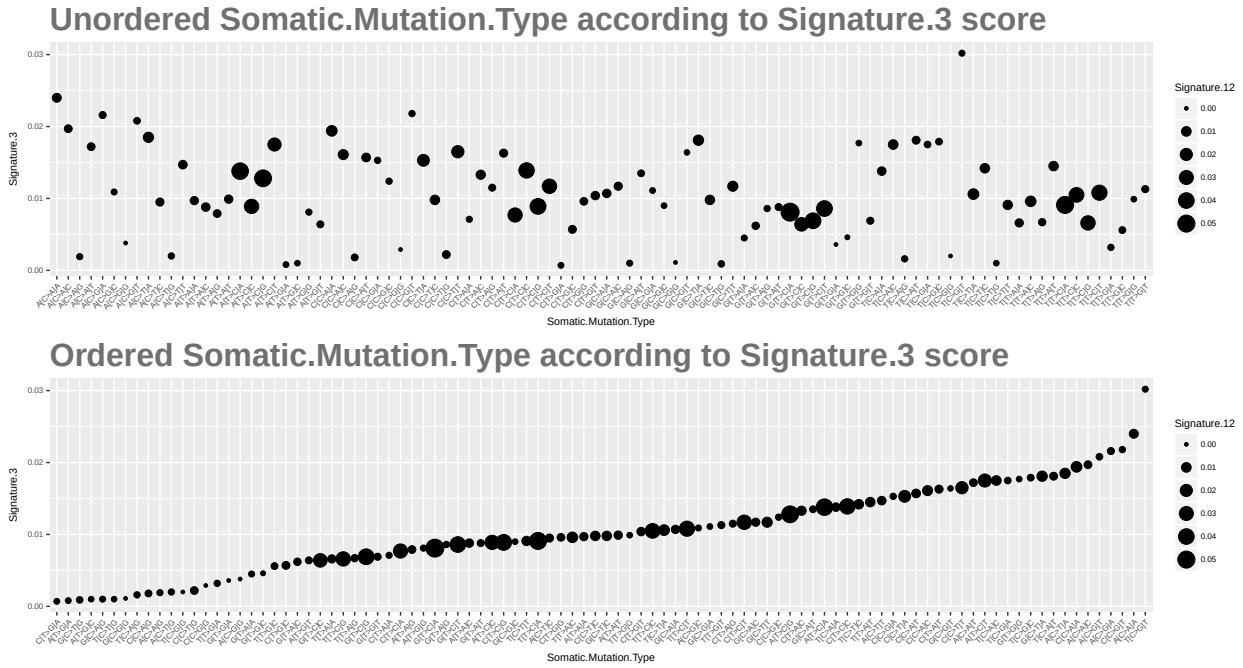


Figure 5: sig3

median_aml

sd_aml

total_aml

mean_breast

median_breast

sd_breast

total_breast

mean_medullo

median_medullo

sd_medullo

total_medullo

mean_pancreas

median_pancreas

sd_pancreas

total_pancreas

A[C>A]A

7.8571429

8

4.4131837

55
70.571429
48
65.041000
8398
24.73
16.5
28.889149
2473
121.866667
98
74.126211
1828
A[C>A]C
4.8571429
3
4.4880794
34
60.949580
39
51.024230
7253
19.70
12.5
24.634910
1970
90.600000
77
53.730545
1359
A[C>A]G
0.8571429
1
0.8997354
6
8.747899

7
6.891678
1041
2.80
2.0
3.305948
280
12.133333
10
9.203002
182
A[C>A]T
5.1428571
4
3.7161168
36
54.873950
33
50.542909
6530
17.72
11.0
23.407277
1772
81.466667
66
60.053389
1222
A[C>G]A
3.8571429
2
3.8047589
27
46.445378
22
45.246912

5527
13.18
10.0
13.096672
1318
64.866667
53
56.628951
973
A[C>G]C
1.8571429
2
1.6761634
13
28.512605
17
24.791287
3393
7.44
7.0
7.455349
744
37.000000
24
28.869163
555
A[C>G]G
0.7142857
0
1.1126973
5
11.159664
6
13.038394
1328
2.37

2.0

2.033333

237

9.466667

8

8.078779

142

A[C>G]T

3.5714286

3

3.1547394

25

47.663865

23

47.465648

5672

10.35

8.0

9.976108

1035

64.066667

32

59.666294

961

A[C>T]A

18.4285714

15

14.3742495

129

78.672269

57

55.095312

9362

31.86

22.5

29.770170

3186

141.933333

133

57.274361

2129

A[C>T]C

8.1428571

5

6.3358391

57

40.663865

33

24.169345

4839

16.07

11.5

14.968826

1607

61.600000

53

29.056103

924

Note. I wrote some stats tables for individual cancer types, then aggregated them (I peaked at how to make a loop here, but ended up performing the task on individual dataset because I couldn't figure out how to avoid "overwriting" at each loop iteration). The summary statistics tables are available here

Some linear regression modeling

I have used this website here discussing broom and variance, and this website here on some broom vignettes. And that's the moment I realized I didn't have two quantitative variables to perform linear regression modelling. This is a recurring problem with me: I always get into performing the analyses before making sure that the dataset format is appropriate (check this hw04 readme file here if you want an example).

So instead of fitting a linear regression of the lifeExp according to year, I will do something slightly different:

Merging and ploting linear regression

I will merge the cancer somatic data with the reference signature data, and I will fit a linear regression between Signature.3 and Signature.12 and then Signature.2 and Signature.13 (see here for more details on mutational signatures).

Note.1. It looks like the T>C somatic mutation group is the one differing the most between Signature.3 and Signature.12, but otherwise, these two signatures could fit "relatively well" a linear regression model

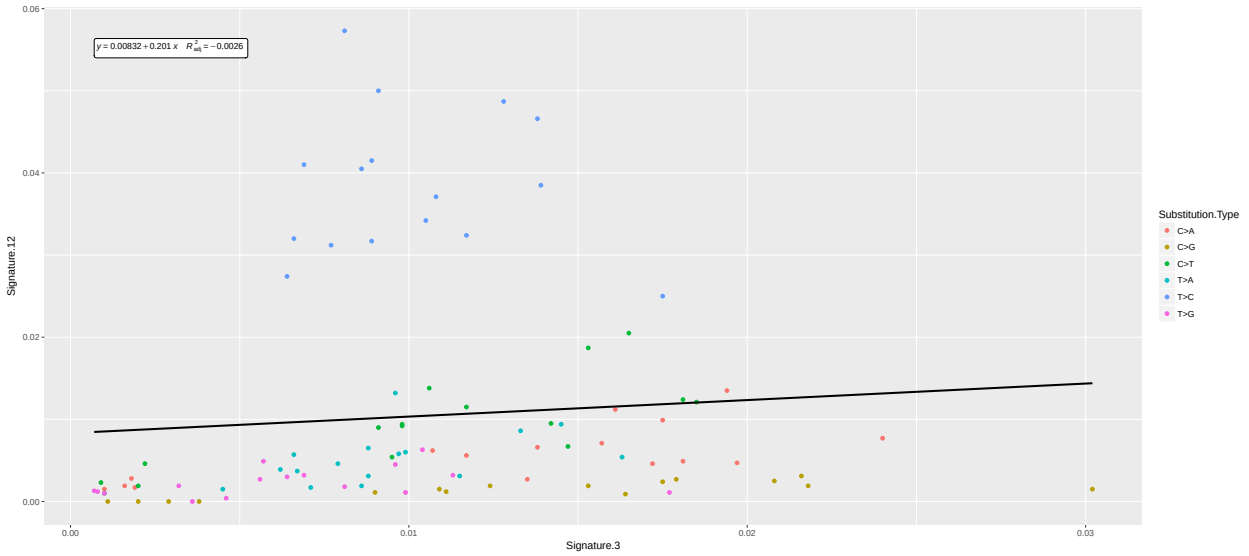


Figure 6: p3_linear_Sig3_Sig12

(although this is obviously not the best model for this type of data, because the adjusted R-squared value is below zero = the model does not fit the data very well).

Note.2. The code and format used above to add a label with the linear regression formula used the following references:

- A stack overflow discussion on Regression equation
- The stack overflow discussion on Label position discussion was also useful.

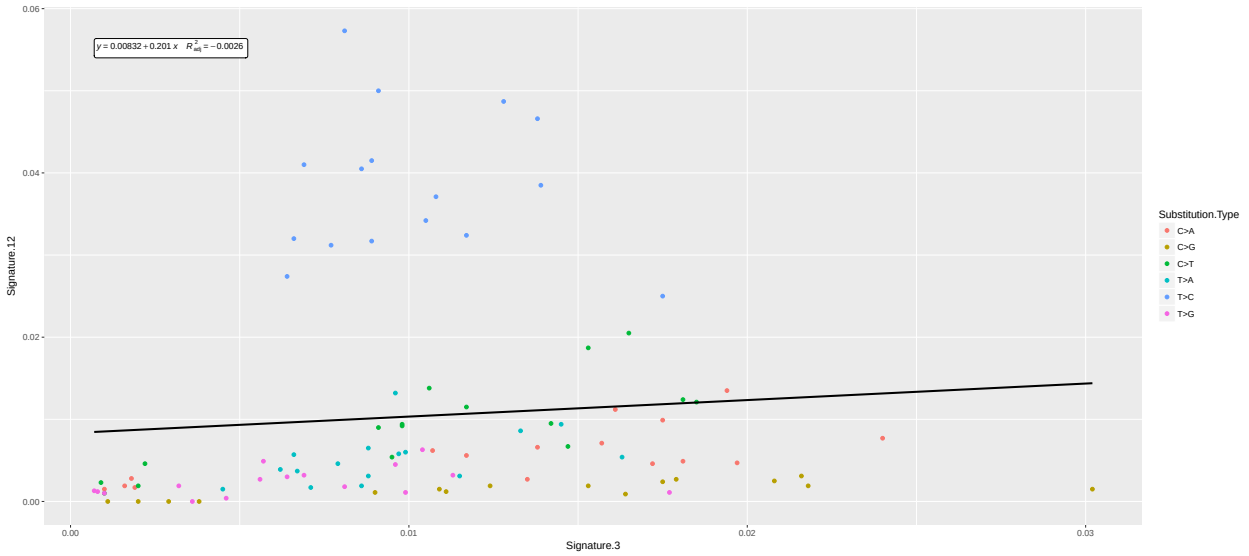
- This article of the cran rstudio website was also used.

```
lm(Signature.12 ~ Signature.3, mut_sig) %>% coef()
```

```
## (Intercept) Signature.3
## 0.008322249 0.201064066
```

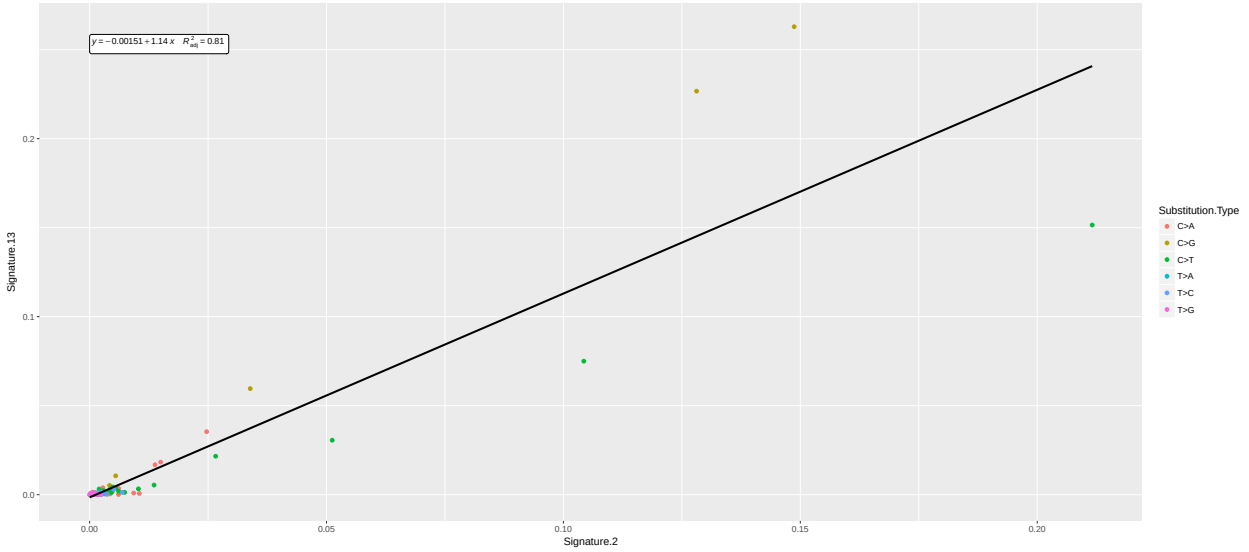
Note. We obtain the same Intercept and Signature.3 coefficient than in our plot (see label formula)

Just for fun, I want to check if a linear regression would be a better fit if I remove the Substitution.Type T>C.



Note. It does seem like the most different values between these two signatures are the T>C Substitution.Type, as our new adjusted R-squared tells us that 15% of the variance of Signature.12 can be explained by Signature.3 if we remove T>C.

Now, let's assess the linear regression model between Signature.2 and Signature.13.



Note. We do get a very high adjusted R-squared value, and this is because these two signatures are thought to be related to the same underlying mutational processes (see here for more details on mutational signatures). Therefore, we can deduce that the Mutation.Type profiles of Signature.2 and Signature.13 are quite similar.

```
lm(Signature.13 ~ Signature.2, mut_sig) %>% coef()
```

```
##  (Intercept)  Signature.2
## -0.001509612  1.144922739
```

I will also show a plot of the number of mutations for all cancer types according to Signature.3 and see what the linear regression looks like.

Find the 3 or 4 "worst" or "best" Signature.3 scores for each cancer_type.

Let's see the cases with the lowest and highest proportion of mutations caused by Signature.3 :

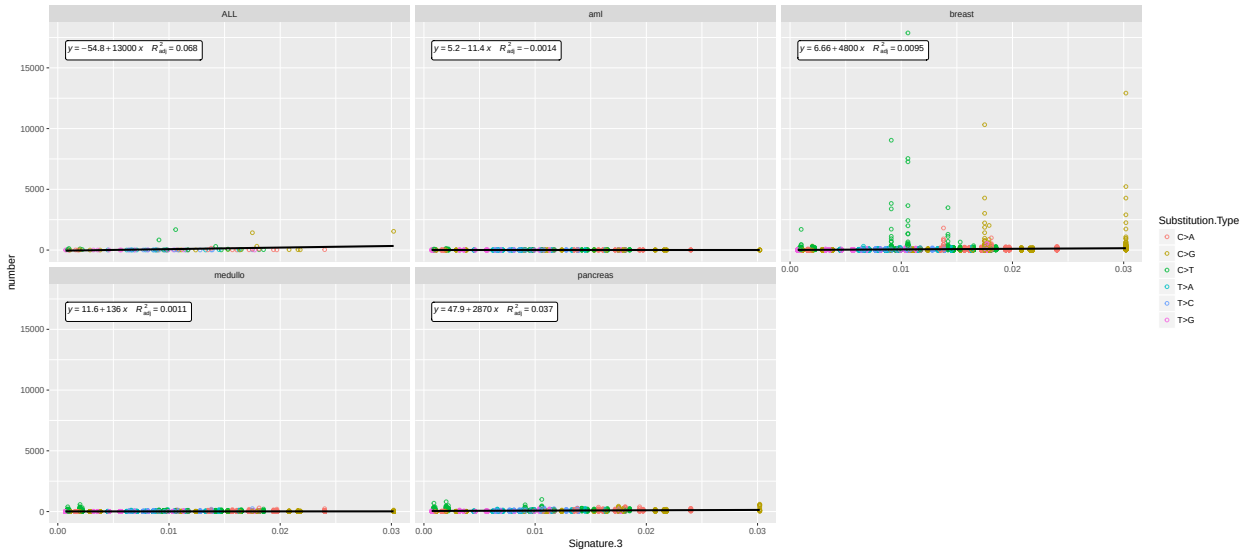


Figure 7: p2_linear_Sig3_all_cancer_mutations

```
all_cancer_types_mut_proportion_signatures  %>%
    group_by(case_id, Signature) %>%
    filter(Signature=="Signature.3") %>%
    arrange(mutations_per_snv_sig_score) %>%
    head(3) %>% kable()
```

case_id

total_number_mutation_per_case

Mutation.Type

number

cancer_type

Substitution.Type

Trinucleotide

Signature

Score

mutations_per_snv_sig_score

proportion_number_each_snv_each_sig

MB28

44

C[T>G]A

0

medullo

T>G

CTA

Signature.3

7e-04

0.0011407

2.59e-05

MB28

44

A[T>G]A

0

medullo

T>G

ATA

Signature.3

8e-04

0.0013037

2.96e-05

MB28

44

G[C>T]G

3

medullo

C>T

GCG

Signature.3

9e-04

0.0014667

3.33e-05

```
all_cancer_types_mut_proportion_signatures  %>%
    group_by(case_id) %>%
    filter(Signature=="Signature.3") %>%
    arrange(desc(mutations_per_snv_sig_score)) %>%
    head(3) %>% kable()
```

case_id

total_number_mutation_per_case

Mutation.Type

number

cancer_type

Substitution.Type

Trinucleotide

Signature

Score

mutations_per_snv_sig_score

proportion_number_each_snv_each_sig

PD4120a

67364

T[C>G]T

12919

breast

C>G

TCT

Signature.3

0.0302

75.34788

0.0011185

PD4120a

67364

A[C>A]A

114

breast

C>A

ACA

Signature.3

0.0240

59.87911

0.0008889

PD4120a

67364

C[C>G]T

225

breast

C>G

CCT

Signature.3

0.0218

54.39019

0.0008074

Note. This actually makes sense because breast cancer usually has higher Signature.3 due to homologous recombination repair deficiency, often secondary to *BRCA1/BRCA2* loss of function.

Generate figures

No worries here, I have generated plenty of figures.

Automate the pipeline (see Makefile)

- mut_clean_genome_text_files.R -> takes as input the file mut_sig_raw.txt and writes the cleaned up/formated version mut_sig.txt
- mut_sig_reorder.R -> reorder the mut_sig.txt Somatic.Mutation.Type according to Signature.3
- mut_sig_tables.R -> uses tidyverse function gather to go from a "wide" to a "long" dataset format and it is also used to perform aggregation taskts, so that the data can be ready for plots later one.
- mut_sig_stat.R -> input the "gathered" dataset files of each cancer type, group by Mutation.Type and outputs files containing summary statistics (mean, median, sd, total)
- mut_sig_plot.R -> input the "gathered" and/or "aggregated" dataset files and produces plots of two types: general data plots, linear modeling plots.
- summary_file.Rmd -> this current file is a summary of this homework, and it has for input the output of the scripts previously mentioned.
- Makefile -> this is the script that runs all the scripts above and coordinate the steps.

Additional material - optional

Here are some samples of the data files I have written:

```
head(all_cancer_types_mut) %>% kable()
```

Mutation.Type

case_id

number

cancer_type

A[C>A]A

PD4020a

33

ALL

A[C>A]C

PD4020a

15

ALL

A[C>A]G

PD4020a

1

ALL

A[C>A]T

PD4020a

24

ALL

A[C>G]A

PD4020a

31

ALL

A[C>G]C

PD4020a

8

ALL

```
head(all_cancer_types_mut_with_ref_sig) %>% kable()
```

Mutation.Type

case_id

number

cancer_type

Substitution.Type

Trinucleotide

Signature.1A

Signature.1B

Signature.2

Signature.3

Signature.4

Signature.5

Signature.6

Signature.7

Signature.8

Signature.9

Signature.10

Signature.11

Signature.12

Signature.13

Signature.14

Signature.15

Signature.16

Signature.17

Signature.18

Signature.19

Signature.20

Signature.21

Signature.R1

Signature.R2

Signature.R3

Signature.U1

Signature.U2

A[C>A]A

PD4020a

33

ALL

C>A

ACA

0.0112

0.0104

0.0105

0.0240

0.0365

0.0149

0.0017

4e-04

0.0368

0.0120

0.0007

2e-04

0.0077

0.0007
0.0001
0.0013
0.0161
0.0018
0.0500
0.0107
0.0013
1e-04
0.0210
0.0137
0.0044
0.0105
0.0221
A[C>A]C
PD4020a
15
ALL
C>A
ACC
0.0092
0.0093
0.0061
0.0197
0.0309
0.0089
0.0028
5e-04
0.0287
0.0067
0.0010
1e-03
0.0047
0.0001
0.0042
0.0040

0.0097
0.0003
0.0076
0.0074
0.0024
7e-04
0.0065
0.0046
0.0047
0.0005
0.0123
A[C>A]G
PD4020a
1
ALL
C>A
ACG
0.0015
0.0016
0.0013
0.0019
0.0183
0.0022
0.0005
0e+00
0.0017
0.0005
0.0003
0e+00
0.0017
0.0001
0.0005
0.0000
0.0022
0.0000
0.0017

0.0005
0.0000
0e+00
0.0000
0.0048
0.0003
0.0000
0.0028
A[C>A]T
PD4020a
24
ALL
C>A
ACT
0.0063
0.0067
0.0037
0.0172
0.0243
0.0092
0.0019
4e-04
0.0300
0.0068
0.0092
2e-04
0.0046
0.0002
0.0296
0.0057
0.0088
0.0032
0.0181
0.0074
0.0029
6e-04

0.0058

0.0081

0.0034

0.0112

0.0118

A[C>G]A

PD4020a

31

ALL

C>G

ACA

0.0018

0.0051

0.0048

0.0216

0.0097

0.0117

0.0013

0e+00

0.0085

0.0048

0.0005

7e-04

0.0031

0.0018

0.0001

0.0011

0.0048

0.0016

0.0014

0.0058

0.0005

5e-04

0.0038

0.0024

0.0012

0.0044
0.0108
A[C>G]C
PD4020a
8
ALL
C>G
ACC
0.0026
0.0043
0.0031
0.0109
0.0054
0.0073
0.0012
0e+00
0.0037
0.0023
0.0003
3e-04
0.0015
0.0014
0.0000
0.0001
0.0024
0.0016
0.0017
0.0019
0.0022
8e-04
0.0046
0.0018
0.0015
0.0065
0.0074

```
head(all_cancer_mutations_per_snv_sig_score) %>% kable()
```

case_id

total_number_mutation_per_case

Mutation.Type

number

cancer_type

Substitution.Type

Trinucleotide

Signature

Score

mutations_per_snv_sig_score

PD4020a

7741

A[C>A]A

33

ALL

C>A

ACA

Signature.1A

0.0112

3.2110815

PD4020a

7741

A[C>A]C

15

ALL

C>A

ACC

Signature.1A

0.0092

2.6376741

PD4020a

7741

A[C>A]G

1

ALL

C>A
ACG
Signature.1A
0.0015
0.4300556
PD4020a
7741
A[C>A]T
24
ALL
C>A
ACT
Signature.1A
0.0063
1.8062333
PD4020a
7741
A[C>G]A
31
ALL
C>G
ACA
Signature.1A
0.0018
0.5160667
PD4020a
7741
A[C>G]C
8
ALL
C>G
ACC
Signature.1A
0.0026
0.7454296