Somatic hypermutations signatures

Anna Obraztsova, Mikhail Shugay

4/24/2017

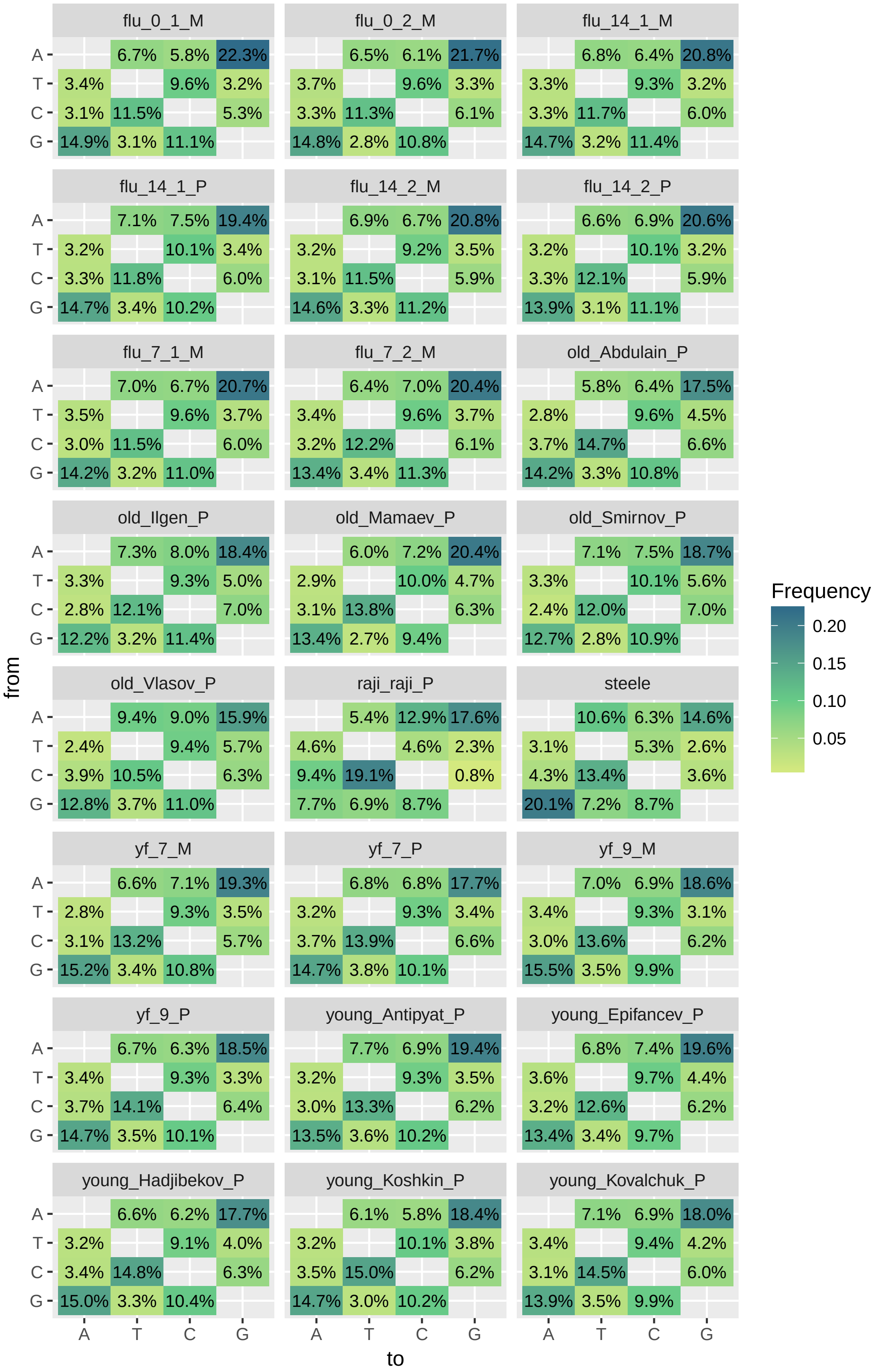
Exploratory data analysis

Load all our data, plot relative substitution frequencies. Here is some summary of what we currently have:

- We have flu (`flu`) and yellow fewer (`YF`) vaccination time-courses which track plasma (`P`) and memory (`M`) B-cells.
- We also have `old` and `young` donors vaccinated against yellow fewer, `P` cells only and no controls unfortunately.
- Raji cell line (`raji`) and data from Steele 2009 (`steele`) are included for reference.

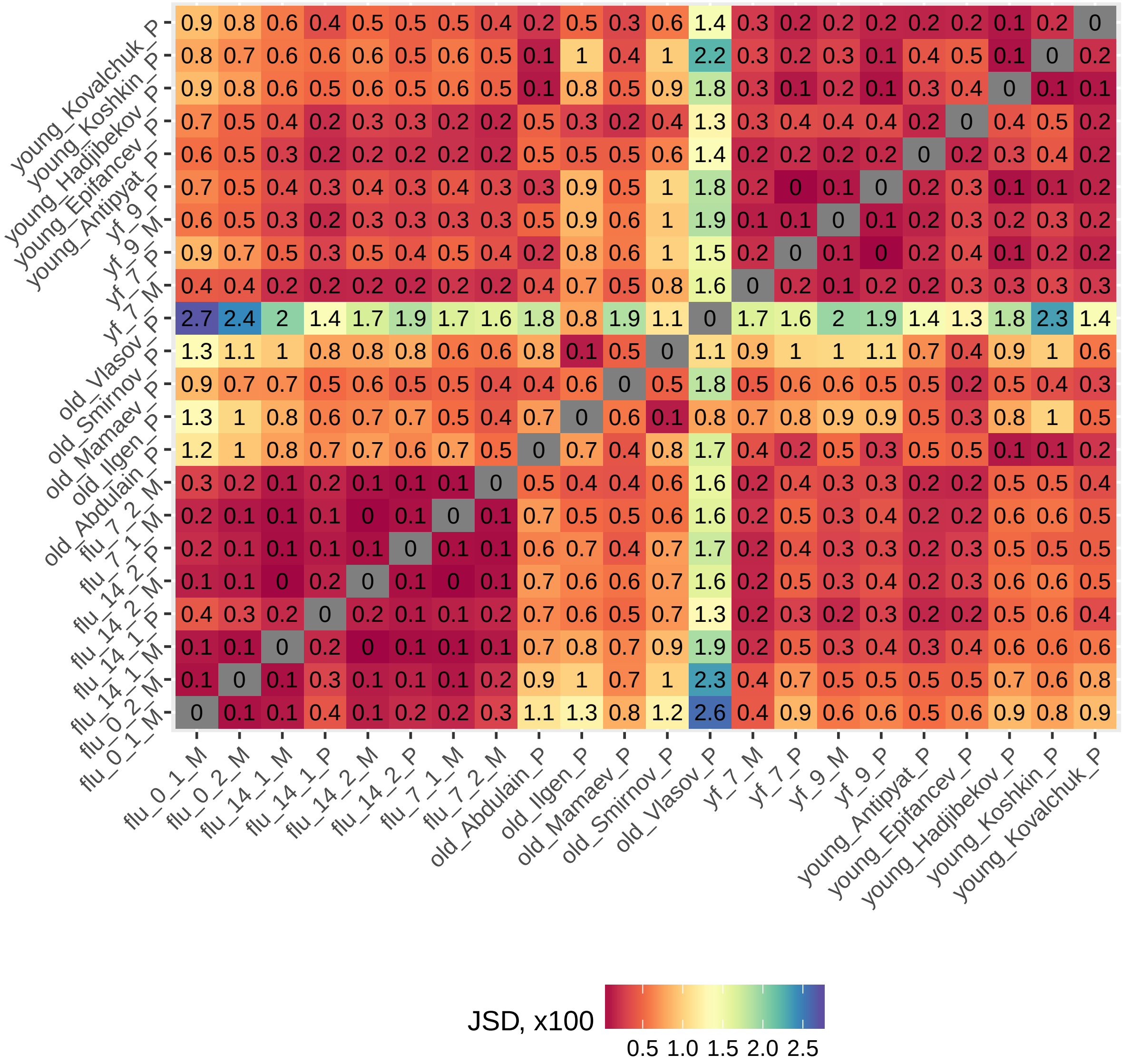
Note that we work with relative fractions of subsitutions which is computed as follows. Let the number of substitutions from base B_i to base B_j be $\#\left(B_i \to B_j\right)$, the absolute substitution frequncy is then $F_{ij} = \#\left(B_i \to B_j\right)/\#B_i$ where $\#B_i$ is the total number of occurences of base B_i in a sample of sequences. The relative frequncy is given by normalizing all F_{ij} to $\sum_{ij} F_{ij} = 1$ (to 100%), i.e. $f_{ij} = F_{ij}/\sum_{lk} F_{lk}$.

Also note that here we ignore abundancies of individual B-cell clonotypes and count each of them only once when summing over substitutions. This is reasonable as it removes substitution frequnecy biases coming from preferential expansion of B-cell clonotypes with certain hypermutations.


flu_0_1_M
6.7% 5.8% 22.3%
3.4% 9.6% 3.2%
3.1% 11.5% 5.3%
14.9% 3.1% 11.1%
flu_0_2_M
6.5% 6.1% 21.7%
3.7% 9.6% 3.3%
3.3% 11.3% 6.1%
14.8% 2.8% 10.8%
flu_14_1_M
6.8% 6.4% 20.8%
3.3% 9.3% 3.2%
3.3% 11.7% 6.0%
14.7% 3.2% 11.4%
flu_14_1_P
7.1% 7.5% 19.4%
3.2% 10.1% 3.4%
3.3% 11.8% 6.0%
14.7% 3.4% 10.2%
flu_14_2_M
6.9% 6.7% 20.8%
3.2% 9.2% 3.5%
3.1% 11.5% 5.9%
14.6% 3.3% 11.2%
flu_14_2_P
6.6% 6.9% 20.6%
3.2% 10.1% 3.2%
3.3% 12.1% 5.9%
13.9% 3.1% 11.1%
flu_7_1_M
7.0% 6.7% 20.7%
3.5% 9.6% 3.7%
3.0% 11.5% 6.0%
14.2% 3.2% 11.0%
flu_7_2_M
6.4% 7.0% 20.4%
3.4% 9.6% 3.7%
3.2% 12.2% 6.1%
13.4% 3.4% 11.3%
old_Abdulain_P
5.8% 6.4% 17.5%
2.8% 9.6% 4.5%
3.7% 14.7% 6.6%
14.2% 3.3% 10.8%
old_Ilgen_P
7.3% 8.0% 18.4%
3.3% 9.3% 5.0%
2.8% 12.1% 7.0%
12.2% 3.2% 11.4%
old_Mamaev_P
6.0% 7.2% 20.4%
2.9% 10.0% 4.7%
3.1% 13.8% 6.3%
13.4% 2.7% 9.4%
old_Smirnov_P
7.1% 7.5% 18.7%
3.3% 10.1% 5.6%
2.4% 12.0% 7.0%
12.7% 2.8% 10.9%
old_Vlasov_P
9.4% 9.0% 15.9%
2.4% 9.4% 5.7%
3.9% 10.5% 6.3%
12.8% 3.7% 11.0%
raji_raji_P
5.4% 12.9% 17.6%
4.6% 4.6% 2.3%
9.4% 19.1% 0.8%
7.7% 6.9% 8.7%
steele
10.6% 6.3% 14.6%
3.1% 5.3% 2.6%
4.3% 13.4% 3.6%
20.1% 7.2% 8.7%
yf_7_M
6.6% 7.1% 19.3%
2.8% 9.3% 3.5%
3.1% 13.2% 5.7%
15.2% 3.4% 10.8%
yf_7_P
6.8% 6.8% 17.7%
3.2% 9.3% 3.4%
3.7% 13.9% 6.6%
14.7% 3.8% 10.1%
yf_9_M
7.0% 6.9% 18.6%
3.4% 9.3% 3.1%
3.0% 13.6% 6.2%
15.5% 3.5% 9.9%
yf_9_P
6.7% 6.3% 18.5%
3.4% 9.3% 3.3%
3.7% 14.1% 6.4%
14.7% 3.5% 10.1%
young_Antipyat_P
7.7% 6.9% 19.4%
3.2% 9.3% 3.5%
3.0% 13.3% 6.2%
13.5% 3.6% 10.2%
young_Epifancev_P
6.8% 7.4% 19.6%
3.6% 9.7% 4.4%
3.2% 12.6% 6.2%
13.4% 3.4% 9.7%
young_Hadjibekov_P
6.6% 6.2% 17.7%
3.2% 9.1% 4.0%
3.4% 14.8% 6.3%
15.0% 3.3% 10.4%
young_Koshkin_P
6.1% 5.8% 18.4%
3.2% 10.1% 3.8%
3.5% 15.0% 6.2%
14.7% 3.0% 10.2%
young_Kovalchuk_P
7.1% 6.9% 18.0%
3.4% 9.4% 4.2%
3.1% 14.5% 6.0%
13.9% 3.5% 9.9%
A
T
C
G
from
A T C G
to
Frequency
0.20
0.15
0.10
0.05

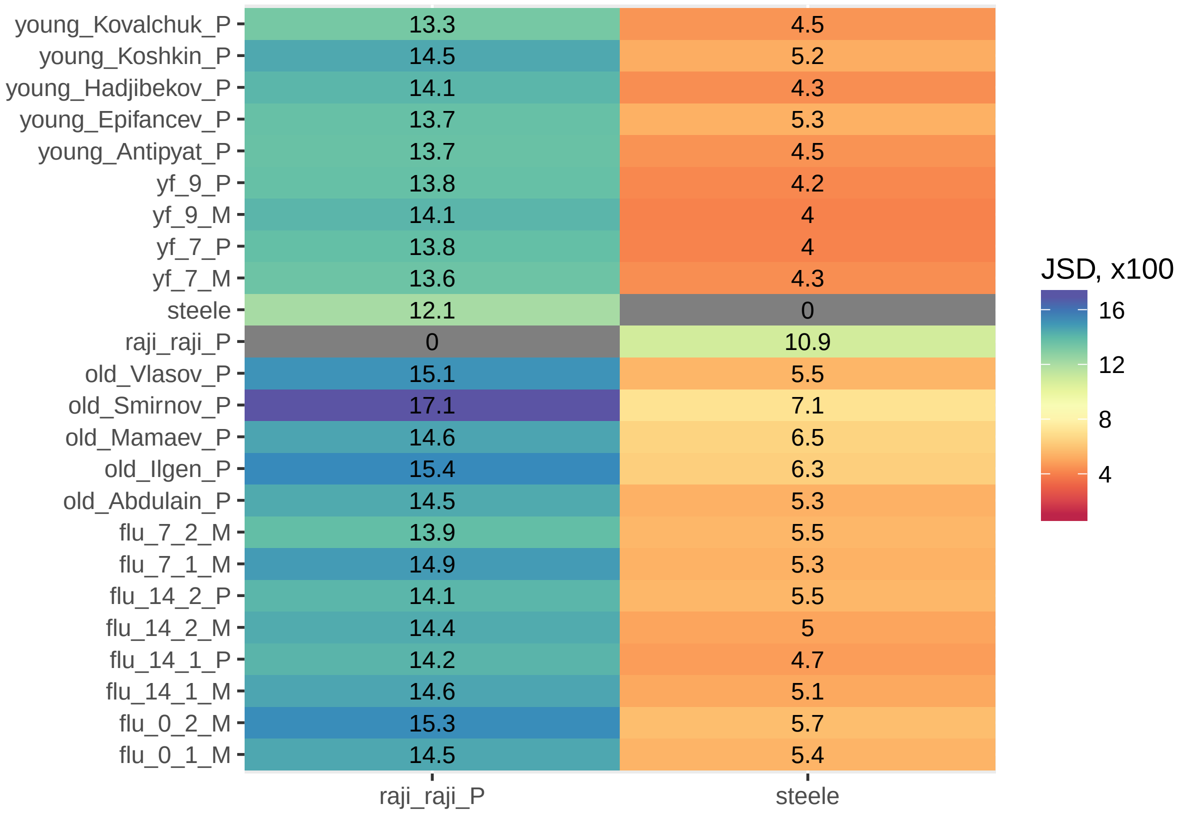
We compute Jensen-Shannon divergences (`JSD`), a metric that is commonly used to compare frequency distributions. The smaller the divergence, the closer are substitution frequency distributions. Of note:

- Old donors appear to be outliers at this plot, but we cannot rule out batch effect in the absence of controls
- Samples for `flu` are highly correlated. Unfortunately this also includes control. All theese samples come from the same donor.

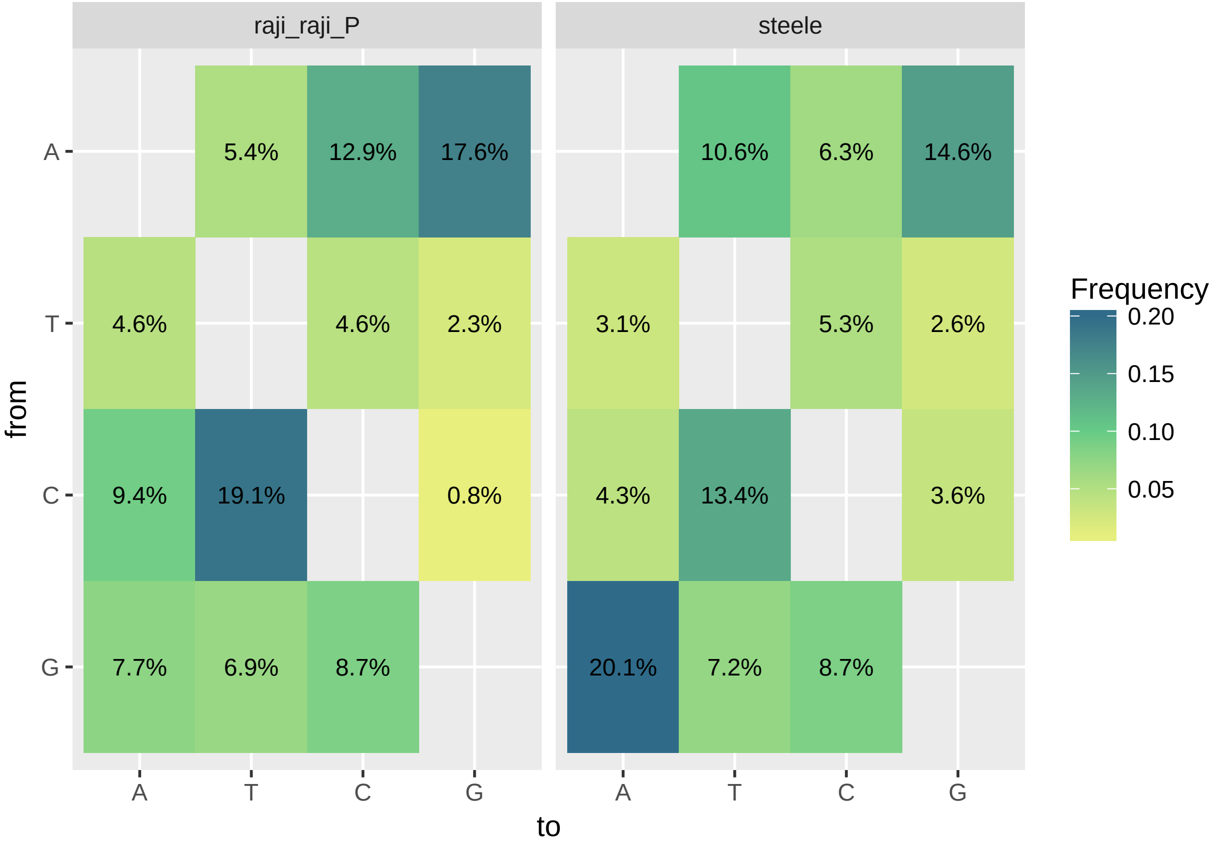


	flu_0_1_M	flu_0_2_M	flu_14_1_M	flu_14_1_P	flu_14_2_M	flu_14_2_P	flu_7_1_M	flu_7_2_M	old_Abdulain_P	old_Ilgen_P	old_Mamaev_P	old_Smirnov_P	old_Vlasov_P	yf_7_M	yf_7_P	yf_9_M	yf_9_P	young_Antipyat_P	young_Epifancev_P	young_Hadjibekov_P	young_Koshkin_P	young_Kovalchuk_P
young_Kovalchuk_P	0.9	0.8	0.6	0.4	0.5	0.5	0.5	0.4	0.2	0.5	0.3	0.6	1.4	0.3	0.2	0.2	0.2	0.2	0.2	0.1	0.2	0
young_Koshkin_P	0.8	0.7	0.6	0.6	0.6	0.5	0.6	0.5	0.1	1	0.4	1	2.2	0.3	0.2	0.3	0.1	0.4	0.5	0.1	0	0.2
young_Hadjibekov_P	0.9	0.8	0.6	0.5	0.6	0.5	0.6	0.5	0.1	0.8	0.5	0.9	1.8	0.3	0.1	0.2	0.1	0.3	0.4	0	0.1	0.1
young_Epifancev_P	0.7	0.5	0.4	0.2	0.3	0.3	0.2	0.2	0.5	0.3	0.2	0.4	1.3	0.3	0.4	0.4	0.4	0.2	0	0.4	0.5	0.2
young_Antipyat_P	0.6	0.5	0.3	0.2	0.2	0.2	0.2	0.2	0.5	0.5	0.5	0.6	1.4	0.2	0.2	0.2	0.2	0	0.2	0.3	0.4	0.2
yf_9_P	0.7	0.5	0.4	0.3	0.4	0.3	0.4	0.3	0.3	0.9	0.5	1	1.8	0.2	0	0.1	0	0.2	0.3	0.1	0.1	0.2
yf_9_M	0.6	0.5	0.3	0.2	0.3	0.3	0.3	0.3	0.5	0.9	0.6	1	1.9	0.1	0.1	0	0.1	0.2	0.3	0.2	0.3	0.2
yf_7_P	0.9	0.7	0.5	0.3	0.5	0.4	0.5	0.4	0.2	0.8	0.6	1	1.5	0.2	0	0.1	0	0.2	0.4	0.1	0.2	0.2
yf_7_M	0.4	0.4	0.2	0.2	0.2	0.2	0.2	0.2	0.4	0.7	0.5	0.8	1.6	0	0.2	0.1	0.2	0.2	0.3	0.3	0.3	0.3
old_Vlasov_P	2.7	2.4	2	1.4	1.7	1.9	1.7	1.6	1.8	0.8	1.9	1.1	0	1.7	1.6	2	1.9	1.4	1.3	1.8	2.3	1.4
old_Smirnov_P	1.3	1.1	1	0.8	0.8	0.8	0.6	0.6	0.8	0.1	0.5	0	1.1	0.9	1	1	1.1	0.7	0.4	0.9	1	0.6
old_Mamaev_P	0.9	0.7	0.7	0.5	0.6	0.5	0.5	0.4	0.4	0.6	0	0.5	1.8	0.5	0.6	0.6	0.5	0.5	0.2	0.5	0.4	0.3
old_Ilgen_P	1.3	1	0.8	0.6	0.7	0.7	0.5	0.4	0.7	0	0.6	0.1	0.8	0.7	0.8	0.9	0.9	0.5	0.3	0.8	1	0.5
old_Abdulain_P	1.2	1	0.8	0.7	0.7	0.6	0.7	0.5	0	0.7	0.4	0.8	1.7	0.4	0.2	0.5	0.3	0.5	0.5	0.1	0.1	0.2
flu_7_2_M	0.3	0.2	0.1	0.2	0.1	0.1	0.1	0	0.5	0.4	0.4	0.6	1.6	0.2	0.4	0.3	0.3	0.2	0.2	0.5	0.5	0.4
flu_7_1_M	0.2	0.1	0.1	0.1	0	0.1	0	0.1	0.7	0.5	0.5	0.6	1.6	0.2	0.5	0.3	0.4	0.2	0.2	0.6	0.6	0.5
flu_14_2_P	0.2	0.1	0.1	0.1	0.1	0	0.1	0.1	0.6	0.7	0.4	0.7	1.7	0.2	0.4	0.3	0.3	0.2	0.3	0.5	0.5	0.5
flu_14_2_M	0.1	0.1	0	0.2	0	0.1	0	0.1	0.7	0.6	0.6	0.7	1.6	0.2	0.5	0.3	0.4	0.2	0.3	0.6	0.6	0.5
flu_14_1_P	0.4	0.3	0.2	0	0.2	0.1	0.1	0.2	0.7	0.6	0.5	0.7	1.3	0.2	0.3	0.2	0.3	0.2	0.2	0.5	0.6	0.4
flu_14_1_M	0.1	0.1	0	0.2	0	0.1	0.1	0.1	0.7	0.8	0.7	0.9	1.9	0.2	0.5	0.3	0.4	0.3	0.4	0.6	0.6	0.6
flu_0_2_M	0.1	0	0.1	0.3	0.1	0.1	0.1	0.2	0.9	1	0.7	1	2.3	0.4	0.7	0.5	0.5	0.5	0.5	0.7	0.6	0.8
flu_0_1_M	0	0.1	0.1	0.4	0.1	0.2	0.2	0.3	1.1	1.3	0.8	1.2	2.6	0.4	0.9	0.6	0.6	0.5	0.6	0.9	0.8	0.9

Compare substitution frequency distributions of our samples with `steele` reference and `raji` cell line. Note that `raji` is an extreme outlier, this is quite obvious from the substitution frequency matrices given above. The data from `steele` is far more similar to our results, but still more than 2 times farther in terms of `JSD` distance from each sample than the sample if from its most distant counterpart in our vaccinated donor set.

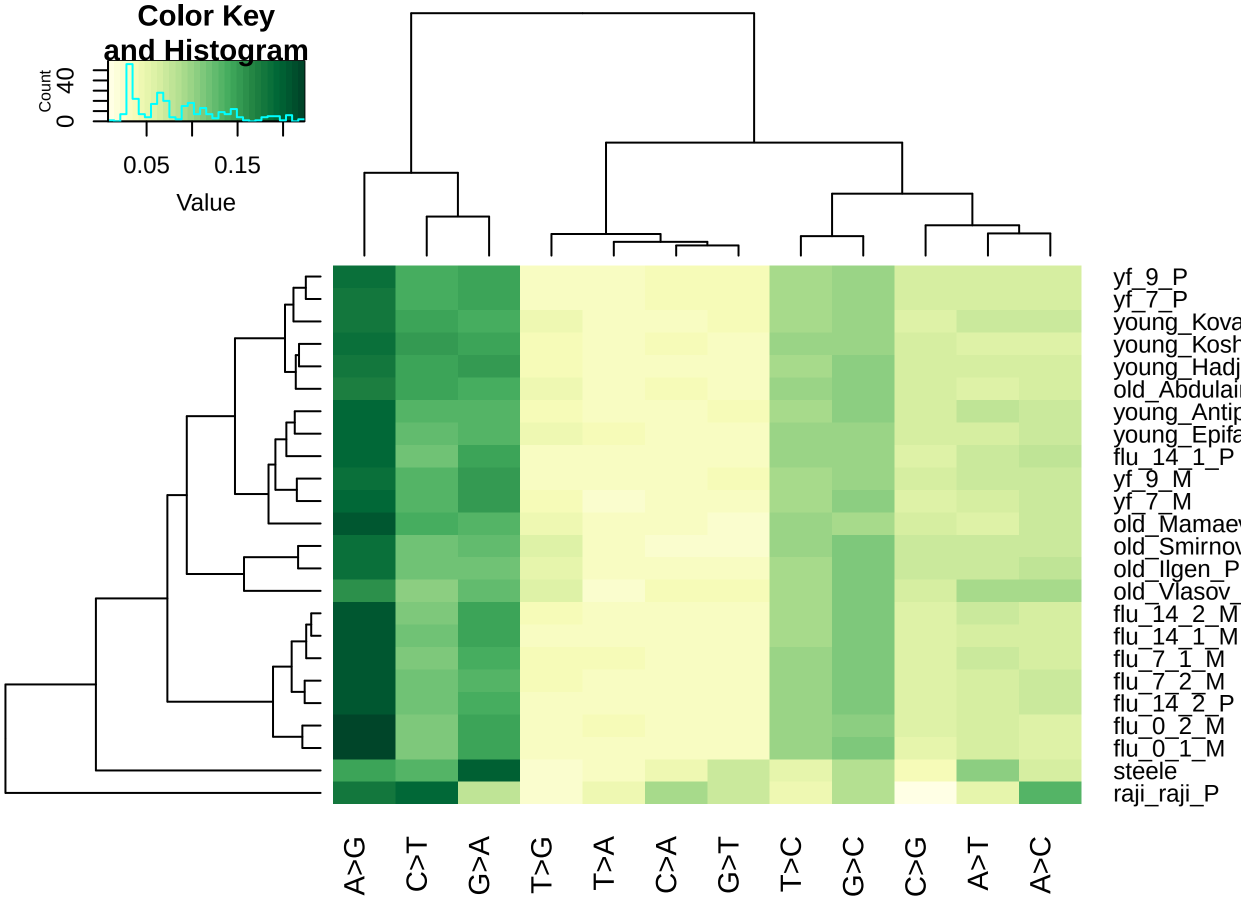


Once more `raji` and `steele` substitution frequencies side-by-side. The `C>>G` rule does not hold for `raji` sample.

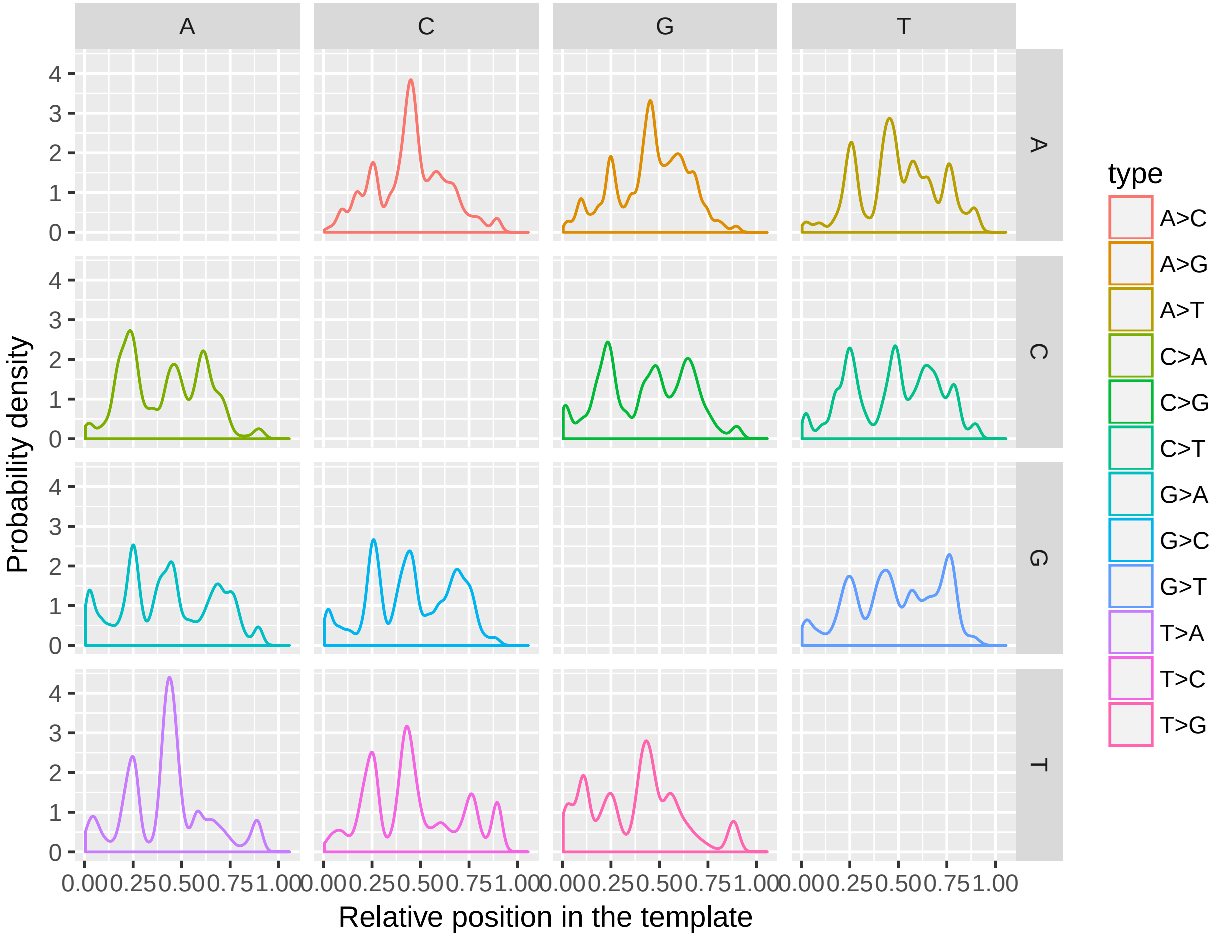


Clustering samples based on mutation profile

```
## Using freq as value column: use value.var to override.
```


Color Key
and Histogram
Count
0
40
0.05
0.15
Value
yf_9_P
yf_7_P
young_Kova
young_Kosh
young_Hadj
old_Abdulai
young_Antip
young_Epifa
flu_14_1_P
yf_9_M
yf_7_M
old_Mamaev
old_Smirnov
old_Ilgen_P
old_Vlasov_
flu_14_2_M
flu_14_1_M
flu_7_1_M
flu_7_2_M
flu_14_2_P
flu_0_2_M
flu_0_1_M
steele
raji_raji_P
A>G
C>T
G>A
T>G
T>A
C>A
G>T
T>C
G>C
C>G
A>T
A>C

Distribution of substitutions by position in template



Mining for ADAR and AID signatures

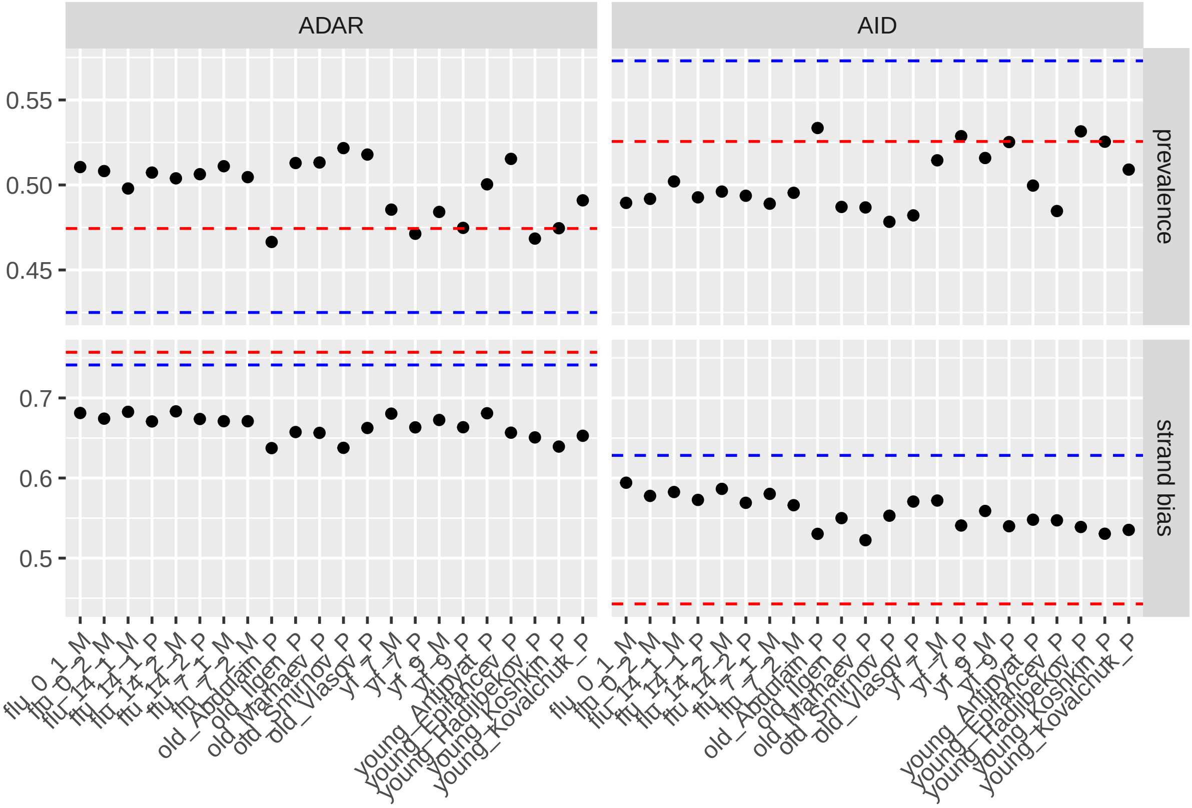
Let us first define a set of four variables corresponding to ADAR/AID signatures:

The fraction of mutations originating from a given base type i is $f_{i \cdot} = \sum_{j \in A,T,G,C} f_{ij}$

- AID prevalence $AID_p = f_{C \cdot} + f_{G \cdot}$
- AID strand bias $AID_s = f_{G \cdot} / (f_{C \cdot} + f_{G \cdot})$
- ADAR prevalence $ADAR_p = f_{A \cdot} + f_{T \cdot}$
- ADAR strand bias $ADAR_s = f_{A \cdot} / (f_{A \cdot} + f_{T \cdot})$

The plot below shows the aforementioned values for each sample. Reference values for `raji` and `steele` are shown in `red` and `blue` respectively.

```
## Using proj, sample, cells, name as id variables
```


ADAR
AID
prevalence
strand bias
0.55
0.50
0.45
0.7
0.6
0.5
flu_0_1_M
flu_0_2_M
flu_14_1_M
flu_14_1_P
flu_14_2_M
flu_14_2_P
flu_7_1_M
flu_7_2_M
old_Abdulaim_P
old_Ilgen_P
old_Mamaev_P
old_Smirnov_P
old_Vlasov_P
yf_7_M
yf_7_P
yf_9_M
yf_9_P
young_Antipyat_P
young_Epifancev_P
young_Hadjibekov_P
young_Koshkin_P
young_Kovalchuk_P
flu_0_1_M
flu_0_2_M
flu_14_1_M
flu_14_1_P
flu_14_2_M
flu_14_2_P
flu_7_1_M
flu_7_2_M
old_Abdulaim_P
old_Ilgen_P
old_Mamaev_P
old_Smirnov_P
old_Vlasov_P
yf_7_M
yf_7_P
yf_9_M
yf_9_P
young_Antipyat_P
young_Epifancev_P
young_Hadjibekov_P
young_Koshkin_P
young_Kovalchuk_P