Ahoj, dávám report ohledně toho, jak pokračuju s prací s tou tabulkou s daty z Boehm článkuMyslím, že se mi konečně povedlo sloučit řádky co měli stejnou aa sekvenci a byli ze stejných organismů (lišily se jen nucleo seq) a vyřadit ty sekvence, kde umi count byl 1 a nevyskytoval se tam víc jak jeden vzorek. Výsledek můžeš vidět na Z:\48\_lab\Project Cell Fate\boehm\_paper\_tcr\250407\_3a\_filtered\_table\_boehmJeště to chci otestovat, jestli se to fakt filtruje správně, protože jsem při každém předchozím testování narazila na nějaký problém a pak to debuggovala hrozně dlouho. Nerada bych totiž špatně vyfiltrovala řádky, co tam mají zůstat nebo naopak tam nechala něco, co není třeba. Z původních 8012 řádků se to při filtru jen na umi count a počet vzorků sníží na 5343, když se k tomu přidá i slučování řádků dle aa sekvence ve stejných vzorcích se to dostane na 4932, což mi přijde docela realistické poměrovně...Mohu Ti svůj postup a detaily buď napsat do zpráv a upravit svůj source code (nyní můžeš najít ve stejné složce pod názvem 250203\_first\_script.R), protože teď to není chronologicky a ne všude jsou komentáře, aby ses na něj mohl podívat, nebo si ohledně toho můžeme zavolat a probrat to spolu. Řekni si, co Ti bude pohodlnější.Jinak doufám, že se máte dobře v Praze

Tuesday, April 8th



ondrejondrej  [12:05 PM](https://labofadaptiveimmunity.slack.com/archives/D064ZN1SNHZ/p1744103117377839)

Ahoj Belo,  
diky. Uz tam ve Finsku konecne mate svetlo.  
Ohledne toho debugingu, moje skills na praci s tabulkami v R jsou ponekud rusty a trvalo by mi dost dlouho nez bych se dostal do stavu, kdy bych mohl kontrolovat Tvuj kod. Mohla bys spis poprosit Juraje, at se na to koukne?  
Kdyz jsem se ted koukal na tu tabulku, tak tam mas sloupec "cell", kde uvadis, z jakeho to je vzorku. Jak resis situaci, ze stejny receptor je v ruznych typech bunek? Je to tak, ze je momentalne na dvou ruznych radcich? To je to, co vlastne potrebujeme nejvice. Dale to UMI proportion je proportion ze vseho nebo z te konkretni sortovane populace (vzorku)? Podle me by to melo byt to druhe. (edited)



BělaBěla  [2:12 PM](https://labofadaptiveimmunity.slack.com/archives/D064ZN1SNHZ/p1744110736368959)

Jasně, můžu požádat Juraje, nicméně mně nejde tolik o debugging, protože teď to funguje dobře, spíš jsem chtěla probrat na základě čeho to filtruju. V té tabulce je 30 sloupců a já jsem to vybrala tak, aby to sloučilo řádky co mají stejnou aa sekvenci a mají stejné tyhle sloupce:  
"genotype" = všechny řádky tam mají "ZGa\_no\_GP", budu ještě muset zjistit, co přesně to znamená  
"cell" = typ buňky dle čeho sortovali, CD4, CD8, DN, DP  
"tissue" = S spleen nebo T thymus  
"tetramer" = n/p  
"raton" = daná myš, např. MM\_3688a  
"Sample.bio.name" = pojmenovaný vzorek dle myši, typu buňky, tkáně, tetrameru např MM\_3688a\_S\_CD4\_Tn  
Jakmile v jednom z těhle sloupců chyběla hodnota, tak jsem to neslučovala ale ani jsem to nevyhazovala z té tabulkyJinak koukám, že excel otvírá zvláštně formát .csv (alespoň moje verze excelu), ale LibreOffice to ukazuje normálně jako tabulku.A to umi.proportion je parametr co tam měli už oni, je to proportion z daného vzorku odpovídající umi.count, stejně jako read.proportion odpovídá read.count z daného vzorku



ondrejondrej  [2:18 PM](https://labofadaptiveimmunity.slack.com/archives/D064ZN1SNHZ/p1744111125508379)

A kdyz jsi slucovala ty radky, tak jsi scitala ten UMI count a read.count v tech radcich?



BělaBěla  [2:25 PM](https://labofadaptiveimmunity.slack.com/archives/D064ZN1SNHZ/p1744111506288259)

Jojo, sčítala jsem umi.count, umi.proportion, read.count a read.proportion, zbytek jsem nechala stejnej jako ten první řádek z těch dvou



ondrejondrej  [2:25 PM](https://labofadaptiveimmunity.slack.com/archives/D064ZN1SNHZ/p1744111551146169)

Tak to je asi v pohode.

[2:27](https://labofadaptiveimmunity.slack.com/archives/D064ZN1SNHZ/p1744111651167779)

ondrej

Pokud v nejakem tom sloupci chyby jasne oznaceni z ceho ten receptor pochazi, tak bych ten radek asi vyhodil. Tim padem, ty .proportion parametry ztrati smysl, a budou se muse spocitat znovu jako pomer sekvenci v ramci konkrektniho vzorku. (edited)



BělaBěla  [2:30 PM](https://labofadaptiveimmunity.slack.com/archives/D064ZN1SNHZ/p1744111853733469)

Takže mám vyhodit všechny řádky, kde chybí hodnota v jednom z těch 6 sloupcí co jsem psala? Je pravda že i po tom filtru na víc jak 1 umi či víc readů s danou aa seq se mi taky poruší proportion, to ještě přepočítám z celkového počtu umi co dostanu po filtrování té tabulky

[2:32](https://labofadaptiveimmunity.slack.com/archives/D064ZN1SNHZ/p1744111946122419)

Běla

A mám tam případně přidat další sloupec, na který se budeme koukat při slučování? Nebo takhle ok?



ondrejondrej  [2:32 PM](https://labofadaptiveimmunity.slack.com/archives/D064ZN1SNHZ/p1744111949906989)

Uplne nevim, jestli absence hodnoty v jakemkoliv sloupci z tech 6. Podle me je nutne se podivat ve kterych ty hodnoty realne chybi a pak si rict, jestli ten TCR je tim neidentifikovatelny nebo jestli ten chybejici udaj byl vlastne redundantni.

[2:34](https://labofadaptiveimmunity.slack.com/archives/D064ZN1SNHZ/p1744112080731819)

ondrej

Zasadni je, aby se pokud tam jsou sekvence se stejnym UMI se to bralo jako jedna sekvence (nejspis oznaceni UMI, ale klidne bych to dal jako COUNT).



BělaBěla  [2:46 PM](https://labofadaptiveimmunity.slack.com/archives/D064ZN1SNHZ/p1744112798051699)

Nejčastěji chybí ta hodnota v tetramer sloupci, což není pro nás tak zásadní, že? Jenže pak někdy chybí ta hodnota ve sloupci cell a to je pro nás důležitý



ondrejondrej  [2:48 PM](https://labofadaptiveimmunity.slack.com/archives/D064ZN1SNHZ/p1744112880903399)

A neni to zjevne z toho sloupce Sample.bio.name?

[2:48](https://labofadaptiveimmunity.slack.com/archives/D064ZN1SNHZ/p1744112897150799)

ondrej

Ja vlastne moc nechapu, jak se mohlo stat, ze tam neco chybi

[2:48](https://labofadaptiveimmunity.slack.com/archives/D064ZN1SNHZ/p1744112904455139)

ondrej

odkud se ten radek pak vzal?



BělaBěla  [2:49 PM](https://labofadaptiveimmunity.slack.com/archives/D064ZN1SNHZ/p1744112970500529)

U těch vzorků kde ty hodnoty chybí je jen "MM\_327\_T" v tom sloupci, což je podle mě jméno té myši a thymus



ondrejondrej  [2:49 PM](https://labofadaptiveimmunity.slack.com/archives/D064ZN1SNHZ/p1744112999345599)

Je mozne, ze je to nejaky experiment, kde to nesortovali?



BělaBěla  [2:50 PM](https://labofadaptiveimmunity.slack.com/archives/D064ZN1SNHZ/p1744113027204149)

Je to víc řádků, namátkově vidím, že je to u téhle myši, kouknu, jestli ještě u jiných taky



ondrejondrej  [2:50 PM](https://labofadaptiveimmunity.slack.com/archives/D064ZN1SNHZ/p1744113045965109)

a koukni, jestli vsechny radky s touhle mysi to maji prazdne

[2:50](https://labofadaptiveimmunity.slack.com/archives/D064ZN1SNHZ/p1744113055849949)

ondrej

pokud ano, tak bych tu mys vyradil celou a je to

[2:51](https://labofadaptiveimmunity.slack.com/archives/D064ZN1SNHZ/p1744113071168069)

ondrej

(pokud nezjistime blize, o co se jedna)



BělaBěla  [3:06 PM](https://labofadaptiveimmunity.slack.com/archives/D064ZN1SNHZ/p1744114006786469)

Jedná se jen o jednu myš a všechny vzorky u ní mají NA v cell a tetramer sloupci

[3:07](https://labofadaptiveimmunity.slack.com/archives/D064ZN1SNHZ/p1744114025224189)

Běla

Tak ji můžeme prostě vyřadit

:+1:1

[3:07](https://labofadaptiveimmunity.slack.com/archives/D064ZN1SNHZ/p1744114043043599)

Běla

A nevím, jestli ji nesortovali, v článku jsem to neviděla

Monday, April 14th