

Statistical analysis for: Viral diversity is an obligate consideration in CRISPR/Cas9 designs for HIV cure.

Methods

The goal of this analysis is to test whether targeting percentages vary by the consensus strain used to create the guides. All analysis was done in R (see code below). To test whether there were overall differences in mean of target percentages, a mixed model was fit with target percentage as the outcome and the consensus strain as the predictors (**lme4** package). A random intercept for each subject by consensus strain was estimated to account for within subject and group variation across the repeated measures. An overall group test (testing whether any mean difference is significantly different than 0) was performed using ANOVA for mixed models using the **lmerTest** package. Post-hoc pairwise comparisons were performed using general linear hypothesis testing for mixed models generated by the **multcomp** and **lsmeans** packages. Adjustment for multiple testing used the Sidak single step method from the **multtest** package (proc = “SidakSS”).

Citations

- **lsmeans**: Lenth RV (2016). “Least-Squares Means: The R Package lsmeans.” *Journal of Statistical Software*, 69(1), 1-33. doi: 10.18637/jss.v069.i01 (URL: <http://doi.org/10.18637/jss.v069.i01>).
- **multcomp**: Hothorn T, Bretz F, Westfall P (2008). “Simultaneous Inference in General Parametric Models.” *Biometrical Journal*, 50(3), 346-363.
- **lme4**: Bates D, Mächler M, Bolker B, Walker S (2015). “Fitting Linear Mixed-Effects Models Using lme4.” *Journal of Statistical Software*, 67(1), 1-48. doi: 10.18637/jss.v067.i01 (URL: <http://doi.org/10.18637/jss.v067.i01>).
- **lmerTest**: Kuznetsova A, Brockhoff PB, Christensen RHB (2017). “lmerTest Package: Tests in Linear Mixed Effects Models.” *Journal of Statistical Software*, 82(13), 1-26. doi: 10.18637/jss.v082.i13 (URL: <http://doi.org/10.18637/jss.v082.i13>).
- **multtest**: Pollard KS, Dudoit S, van der Laan MJ (2005). *Multiple Testing Procedures: R multtest Package and Applications to Genomics, in Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer.

R code

Data read in

```
suppressPackageStartupMessages({
  library(broom)
  library(lmerTest) # mixed models
  library(multcomp) # statistical tests
  library(tidyverse) # data manipulation and plots
  library(knitr)
  library(kableExtra)
  theme_set(theme_bw())
})

opts_chunk$set(tidy = TRUE, cache = TRUE, messages = FALSE, warning = FALSE, echo = FALSE)

dat_all = read_csv(file = "All_pts_matches_highcov.csv") %>%
  group_by(patient, gseq) %>%
  mutate(
    total_consensus = n_distinct(from_consensus),
    groupings = paste(unique(from_consensus), collapse = ","),
    range_pct = diff(range(tgt_match_perc_exact)),
    conserved = total_consensus >= 2
  )

## Parsed with column specification:
## cols(
##   patient = col_character(),
##   from_consensus = col_character(),
##   gseq = col_character(),
##   dir = col_character(),
##   guide = col_character(),
##   pam = col_character(),
##   match_conseq_start = col_integer(),
##   match_conseq_end = col_integer(),
##   tgt_depth = col_integer(),
##   tgt_match_perc_exact = col_double(),
##   has_degenerate_bases = col_logical()
## )
```

Analysis and models

```
summary = dat_all %>% group_by(from_consensus, patient) %>% summarize(n = n(),
  mean_match = mean(tgt_match_perc_exact)) %>% group_by(from_consensus) %>%
  summarize(patients = n_distinct(patient), total_targets = sum(n), mean_pct = paste0(round(100 *
    mean(mean_match), 1), "%"))

### mixed model - pooled con group ###

dat_all$new_var = with(dat_all, ifelse(from_consensus == "patient", "patient",
  "pooled"))
lmm_pool = lmer(tgt_match_perc_exact ~ new_var + (from_consensus | patient),
  data = dat_all)

pooled_results = data.frame(Model = "Pooled", lhs = "Pooled groups - patient")
pooled_results$estimate = coef(summary(lmm_pool))["new_varpooled", "Estimate"]
pooled_results$pvalue_unadjusted = coef(summary(lmm_pool))["new_varpooled",
  "Pr(>|t|)"]

### mixed model - separate by con groups ###

lmm = lmer(tgt_match_perc_exact ~ from_consensus + (from_consensus | patient),
  data = dat_all)

ph_tests = summary(glht(lmm, lsmeans::lsm(pairwise ~ from_consensus)), test = adjusted("none")) %>%
  tidy() %>% select(lhs, estimate, p.value) %>% rename(pvalue_unadjusted = p.value)

# Overall F-test
grp_test = anova(lmm)$`Pr(>F)`

### combine results and adjust p-values ###

all_results = bind_rows(pooled_results, ph_tests %>% mutate(Model = " "))

all_results$pvalue_adjusted = multtest::mt.rawp2adjp(all_results$pvalue_unadjusted,
  proc = "SidakSS")$adjp[, 2]
```

Table 1: Summary of data.

from_consensus	patients	total_targets	mean_pct
A	4	173	86.5%
B	4	488	85.3%
C	4	99	84.7%
M	4	365	85.3%
patient	4	1173	81.6%

Table 2: Mixed model results: post-hoc comparisons (p-values corrected using single-step method). Overall group test (at least one group is different) ($p = 0.148$)

Model	Comparison	Mean diff. (%)	Adj. p-value (unadj.)
	A - B = 0	1.2%	0.458 (0.523)
	A - C = 0	1.71%	0.631 (0.408)
	A - M = 0	1.17%	0.804 (0.502)
	B - C = 0	0.51%	0.997 (0.836)
	B - M = 0	-0.03%	1 (0.978)
	C - M = 0	-0.54%	1 (0.812)
	A - patient = 0	4.88%	0.996 (0.138)
	B - patient = 0	3.68%	1 (0.054)
	C - patient = 0	3.17%	1 (0.388)
	M - patient = 0	3.71%	1 (0.087)
Pooled	Pooled groups - patient = 0	3.43%	0.251 (0.026)

Version information

Table 3: Reproducibility Software Session Information

name	value.V1
version	R version 3.5.0 (2018-04-23)
system	x86_64, darwin15.6.0
ui	X11
language	(EN)
collate	en_US.UTF-8
tz	America/Los_Angeles
date	2018-05-17
repo	https://github.com/proychou/CRISPR
location	https://github.com/proychou/CRISPR
file name	CRISPR_stats.Rmd

Table 4: Reproducibility Software Package Version Information

package	version	date	source
base	3.5.0	2018-04-24	local
bindrcpp	0.2.2	2018-03-29	CRAN (R 3.5.0)
Biobase	2.40.0	2018-05-01	Bioconductor
BiocGenerics	0.26.0	2018-05-01	Bioconductor
broom	0.4.4	2018-03-29	CRAN (R 3.5.0)
data.table	1.11.2	2018-05-08	CRAN (R 3.5.0)
datasets	3.5.0	2018-04-24	local
dplyr	0.7.4	2017-09-28	CRAN (R 3.5.0)
forcats	0.3.0	2018-02-19	CRAN (R 3.5.0)
ggplot2	2.2.1	2016-12-30	CRAN (R 3.5.0)
graphics	3.5.0	2018-04-24	local
grDevices	3.5.0	2018-04-24	local
kableExtra	0.8.0	2018-04-05	CRAN (R 3.5.0)
knitr	1.20	2018-02-20	CRAN (R 3.5.0)
lme4	1.1-17	2018-04-03	CRAN (R 3.5.0)
lmerTest	3.0-1	2018-04-23	CRAN (R 3.5.0)
lsmeans	2.27-62	2018-05-11	CRAN (R 3.5.0)
MASS	7.3-50	2018-04-30	CRAN (R 3.5.0)
Matrix	1.2-14	2018-04-13	CRAN (R 3.5.0)
methods	3.5.0	2018-04-24	local
multcomp	1.4-8	2017-11-08	CRAN (R 3.5.0)
multtest	2.36.0	2018-05-01	Bioconductor
mvtnorm	1.0-7	2018-01-26	CRAN (R 3.5.0)
parallel	3.5.0	2018-04-24	local
purrr	0.2.4	2017-10-18	CRAN (R 3.5.0)
readr	1.1.1	2017-05-16	CRAN (R 3.5.0)
stats	3.5.0	2018-04-24	local
stringr	1.3.1	2018-05-10	CRAN (R 3.5.0)
survival	2.42-3	2018-04-16	CRAN (R 3.5.0)
TH.data	1.0-8	2017-01-23	CRAN (R 3.5.0)
tibble	1.4.2	2018-01-22	CRAN (R 3.5.0)
tidyr	0.8.0	2018-01-29	CRAN (R 3.5.0)
tidyverse	1.2.1	2017-11-14	CRAN (R 3.5.0)
utils	3.5.0	2018-04-24	local