Reproducible Replicates in Peptide Array Data

Tun Lee Ng and Michael A. Newton

April 30, 2020

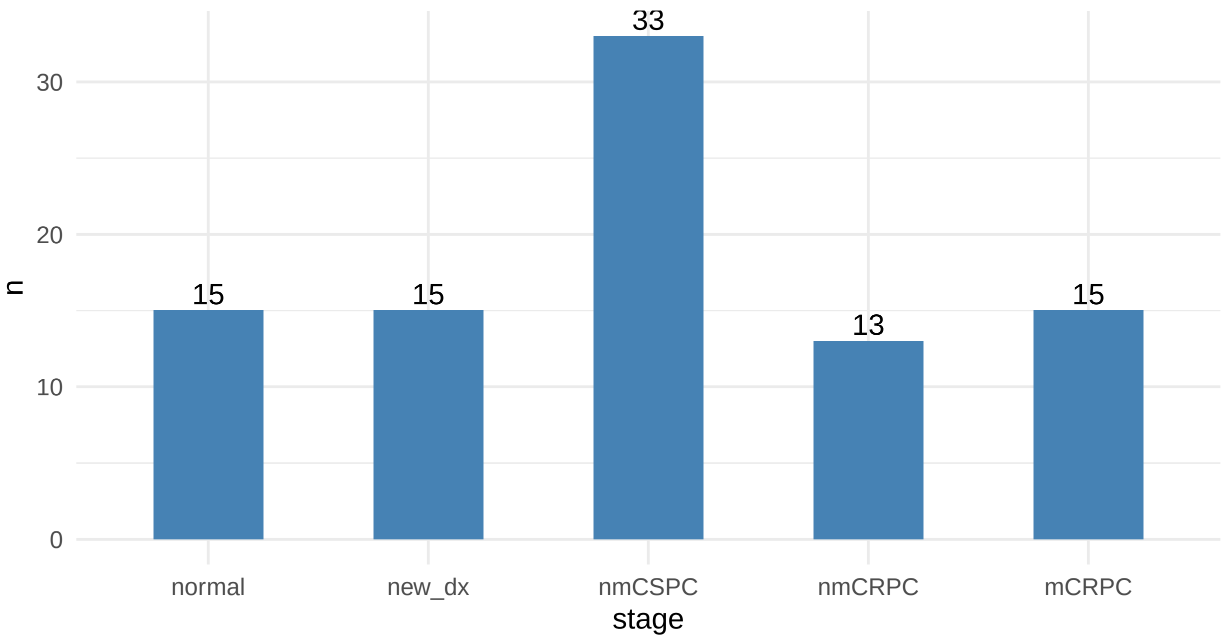
Contents

1 Introduction

This project aims to characterize antibody responses to a wide variety of proteins in prostate cancer patients at different stages of the disease. 16-mer peptides spanning the amino acid sequences of these 1611 gene products, and overlapping by 12 amino acids, were used to generate a microarray comprising 177,604 peptides. In this study, there were healthy subjects and patients with different stages of prostate cancer

- `new_dx`: newly diagnosed,
- `nmCSPC`: non-metastatic castration-sensitive,
- `mCSPC`: metastatic castration-sensitive,
- `nmCRPC`: non-metastatic castration-resistant,
- `mCRPC`: metastatic castration-resistant

Recall that these are not distinct patient counts, because there were 11 patients who were measured at two different stages. We removed these patients' earlier records to ensure unique patient data. Number of replicates for each patient, `rep` could 1, 2, or 3. We remove patients with no technical replicates. So, now we are left with:



Next, we take $\log_2$ transformation of the fluorescence intensity.

2 Assess Reproducibility of Replicates

Hemanth has assessed the issue of replicate reproducibility by looking at correlation coefficients between patients' fluorescence levels. Another approach is to measure how much variation the technical replicates are contributing to the overall variation in the data. If the replicates are reproducible, ie. they "largely agree with one another", then the technical variation in the dataset should be minimal.

Everytime when the fluorescence levels were measured (with replicates) for patient's stage effects, there are two sources of random variation at play, namely

- patient/subject random effect: this reflects the biological variation of a patient (as opposed to the `fixed effect` term, which would be the cancer stage effect in this experiment)

- (residual) random error: measuring replicates of a patient is itself a source of technical variation.

Specifically,

$$y_{ijk} = \mu + \beta_i + b_j + \epsilon_{ijk},$$

where

- y_{ijk} denotes the $\log_2$ fluorescence level of a replicate,

- μ denotes the grand mean/intercept,
- β_i denotes the fixed effect term, ie. cancer stage, with i indexing the patients' cancer stage,

- b_j denotes the random effect term, ie. individual patient, with j indexing the patients,

- ϵ_{ijk} denotes the (residual) random error of the model, with k indexing the replicates.

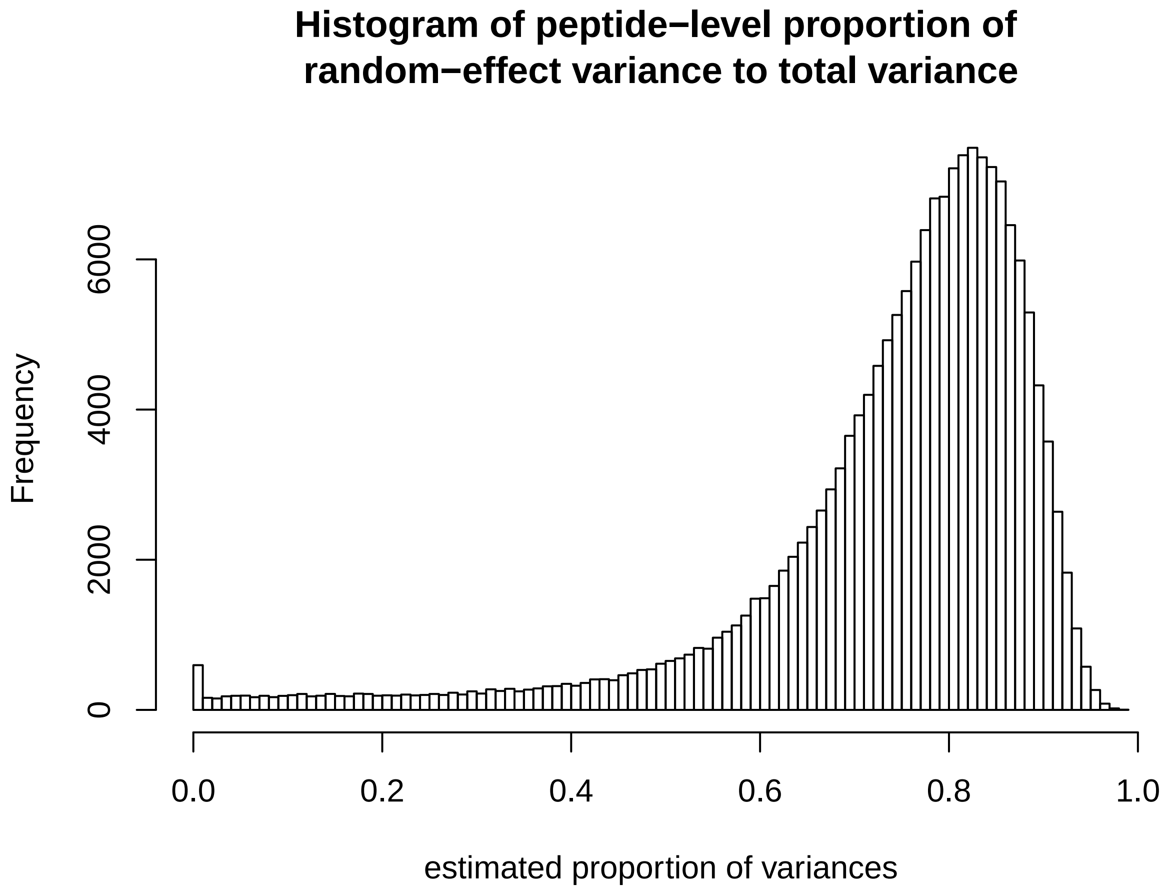
This is the linear mixed-effects model, which we deploy using the R package `lme4` with the following (pseudo)-syntax

$\texttt{lmer}(y \sim \mathrm{stage} + (1 \,|\, \mathrm{patient}))$.

The model estimates the two sources of variation: $\hat{\sigma}_b^2$ (biological variation) and $\hat{\sigma}_\epsilon^2$ (technical variation). Ideally, when the replicates "largely agree with one another", biological variation should dominate technical variation since the replicates' variance $\hat{\sigma}_\epsilon^2$ is minimal. Hence, we are interested in the estimated proportion of random-effect variance to total variance

$$\frac{\hat{\sigma}_b^2}{\hat{\sigma}_b^2 + \hat{\sigma}_\epsilon^2},$$

and we would like to see this ratio to be close to one. For each of the 177k peptides, we deploy this mixed-effect model, and plot the histogram of the estimated proportions of variances.



As expected, the histogram amasses at values near one, indicating that most of the variation in the ($\log_2$) fluorescence data is attributable to the biological variation of the patients and not the technical replicates themselves, which also suggests reproducibility of the replicates. The little spike at zero estimated proportions is due to the 479 singular cases where the fitted random-effect variance $\hat{\sigma}_b^2$ is close/equal to zero.