Using [general linear] models to adjust for total cell number and technical artefacts in high-throughput screens

Stanley E. Lazic[1,*] Paul Selzer[1] Sebastian Hoersch[1]
Marjo Goette[1] Christine Halleux[1]

[1]Novartis Institutes for Biomedical Research, Basel, Switzerland

*Current address: Quantitative Biology, AstraZeneca, Cambridge, UK

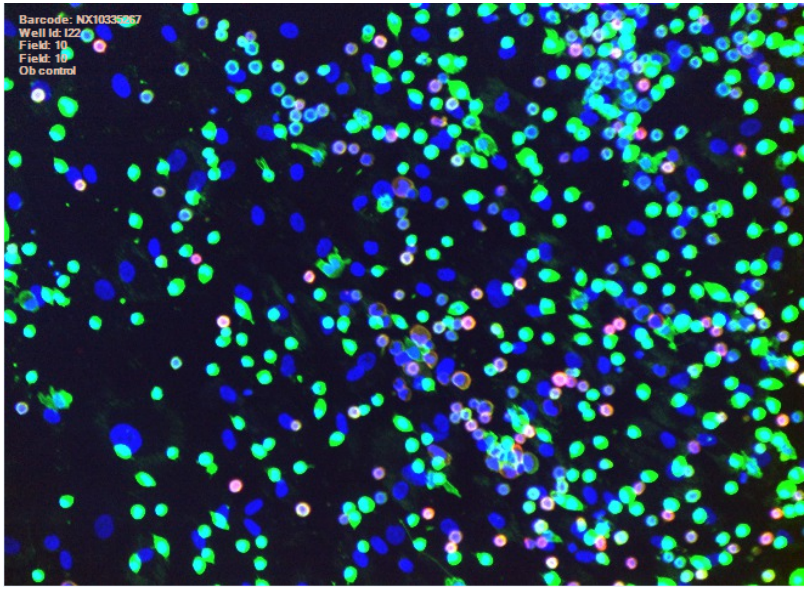
5 Oct 2016

AstraZeneca

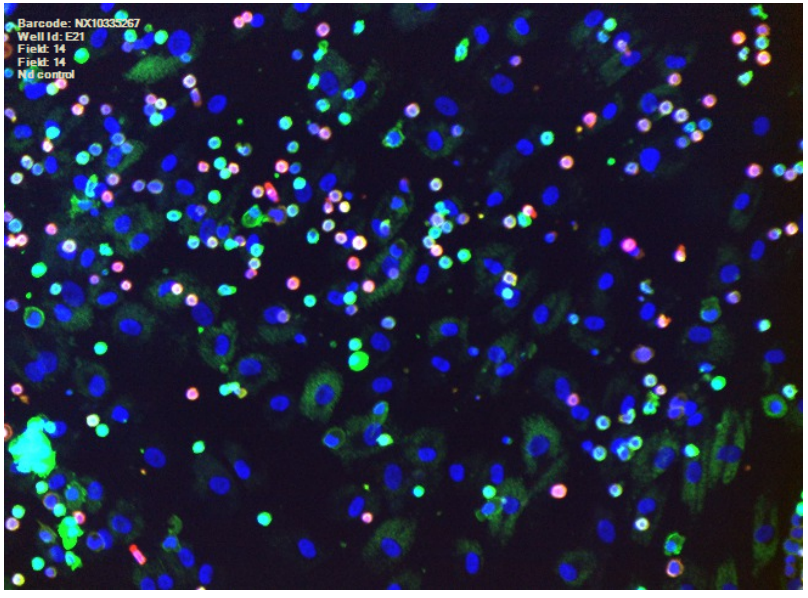
NOVARTIS

High content images

Positive Control



Negative Control



Blue = Nuclear staining

Green = Total cell count

Red = Positive cells

Preprocessing

- **Aggregating/summarising:** Data reduction (e.g. mean).
- **Normalising/standardising:** Making the data “the same” (e.g. z-scores).
- **Correcting/adjusting:** Removing known sources of bias or variation (e.g. subtracting baseline, dividing by body weight).
- **Transforming:** Application of a function to one variable and the same function is applied to all elements in that variable (e.g. log, sqrt).
- **Filtering:** Removing data (e.g. outliers, bad samples, whole variables).

STATISTICS IN MEDICINE
Statist. Med. 2009; **28**:3189–3209
Published online 19 May 2009 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3603

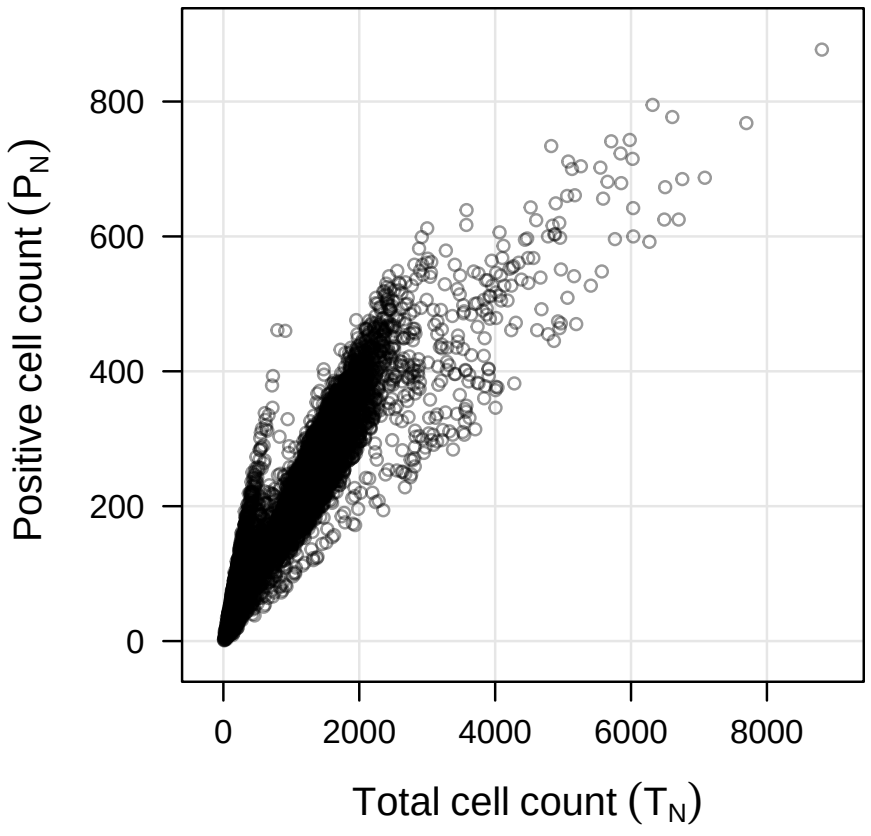
Measurement in clinical trials: A neglected issue for statisticians?

Stephen Senn[1, *, †] and Steven Julious[2]

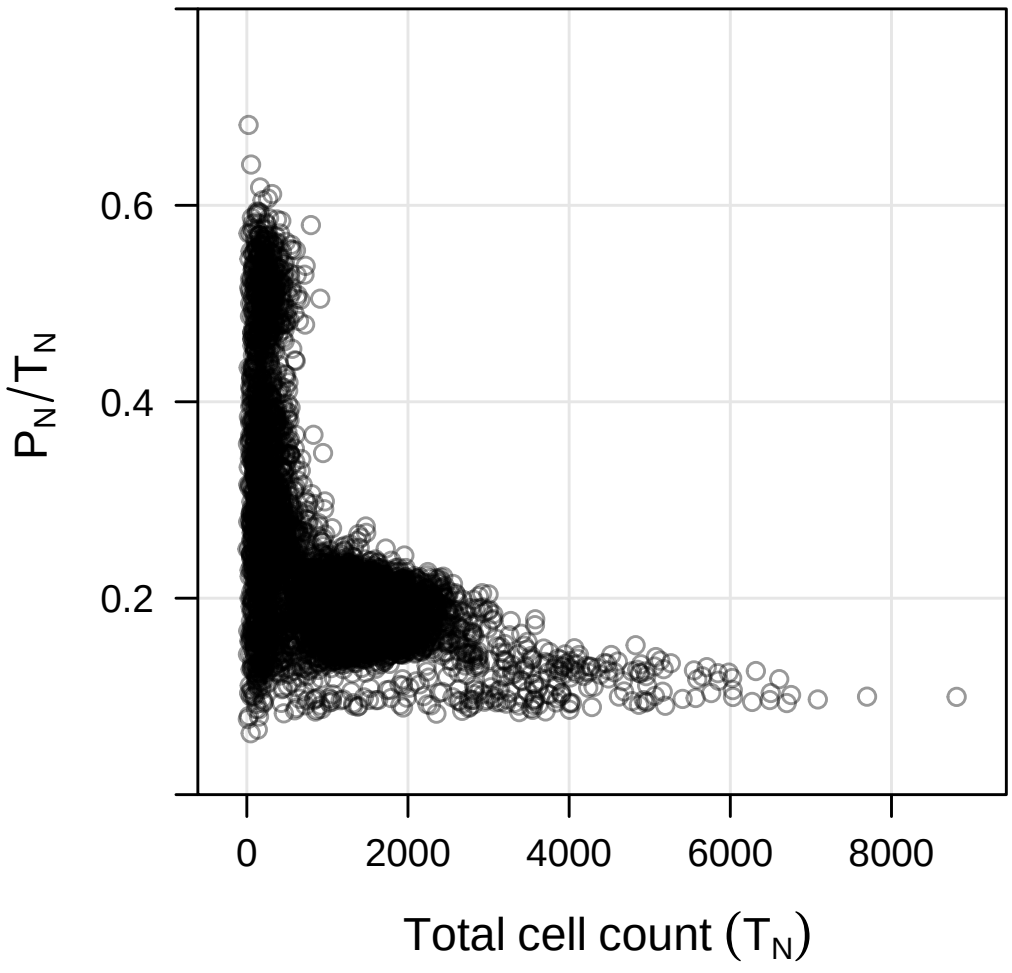
[1]*Department of Statistics, University of Glasgow, Glasgow G12 9LL, U.K.*
[2]*University of Sheffield, Sheffield, U.K.*

- Crude corrections
- Correcting for post-randomisation covariates

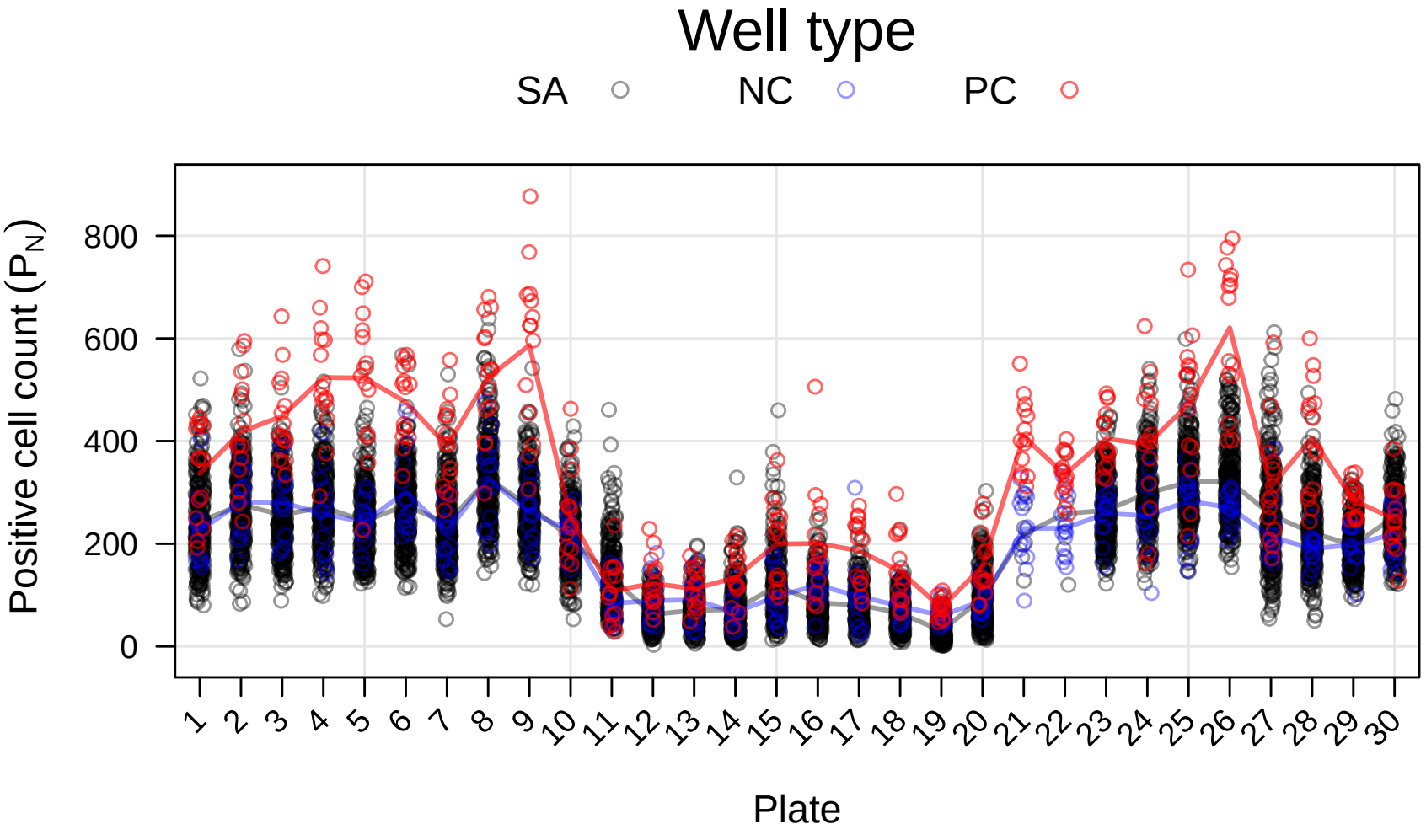
Number of positive cells depends on total cell count



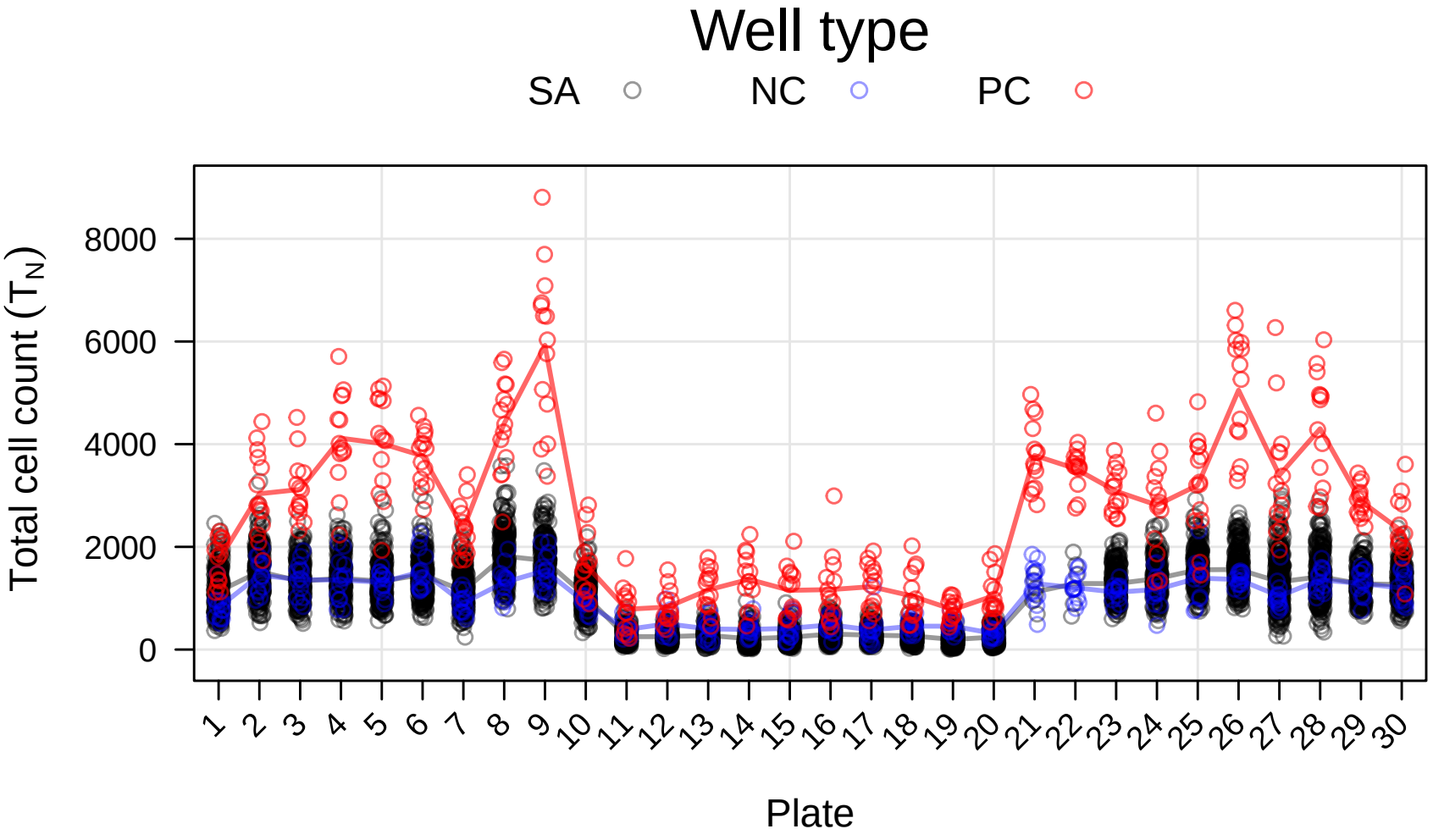
Dividing by total cell count doesn't work



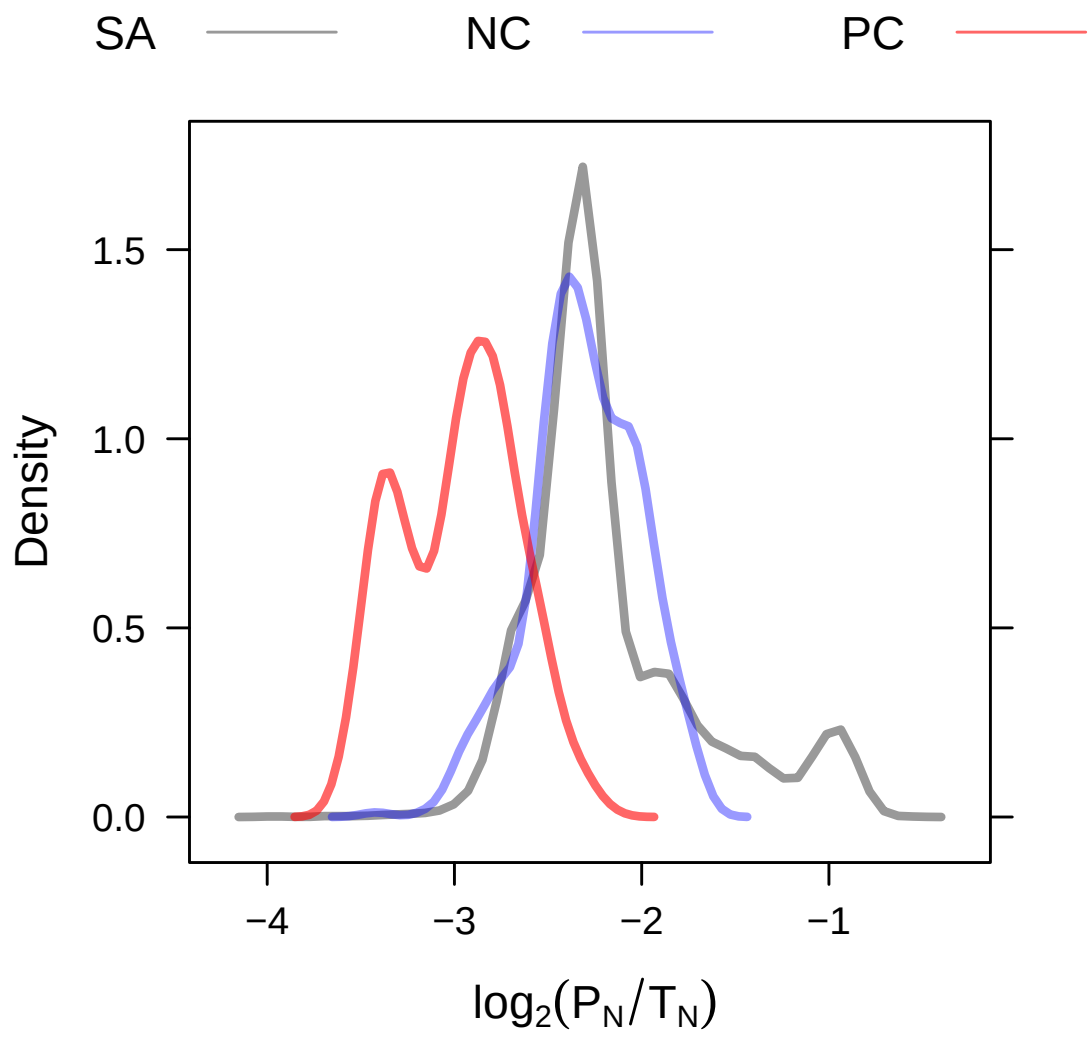
Batch and plate effects are present



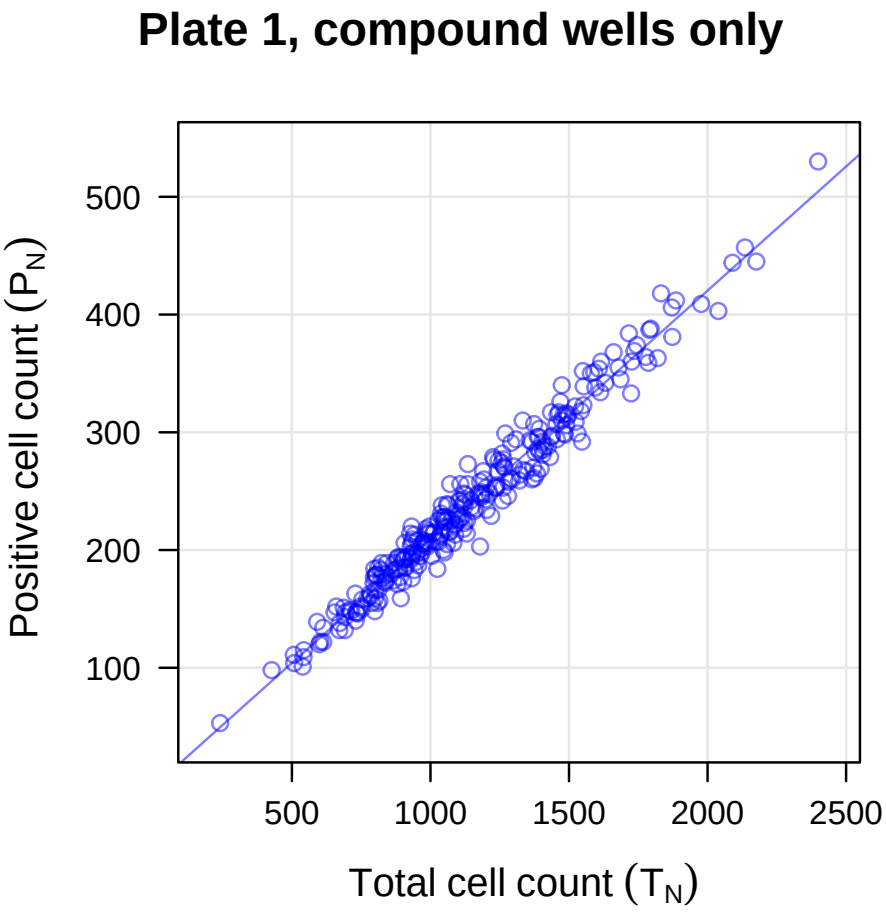
Batch and plate effects are present



Positive controls < negative controls(!)



What do you want to find?



- Largest P_N?
- Largest positive residual?
- Largest positive residual for $T_N < N$?

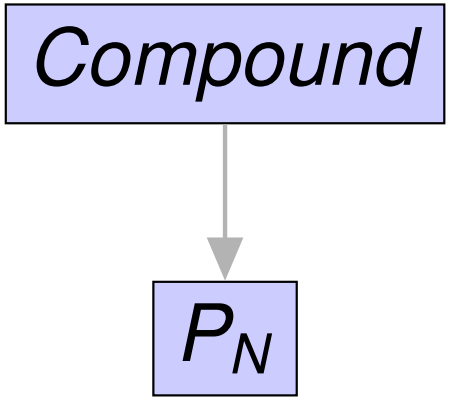
Summary of effects

- P_N affected by T_N.
- P_N and T_N affected by batches and plates.
- P_N and T_N affected by well type.

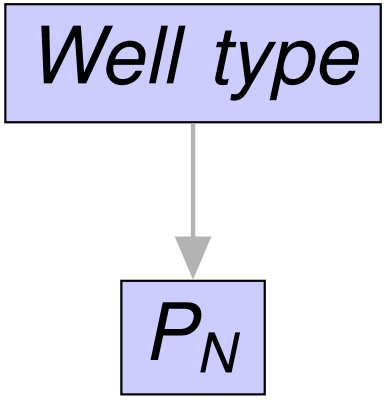
How to remove the effects?

- $P_N / T_N \rightarrow$ performs poorly.
- Standardise plates (mean/SD or median/MAD) $\rightarrow$ division by small numbers.
- What order: standardise ratios or standardise and then calculate ratios?

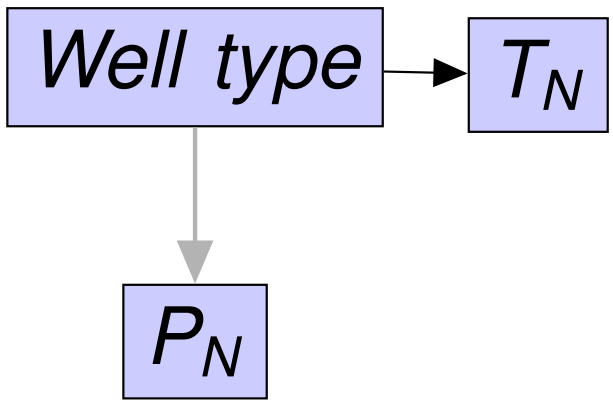
Graphical representation of effects



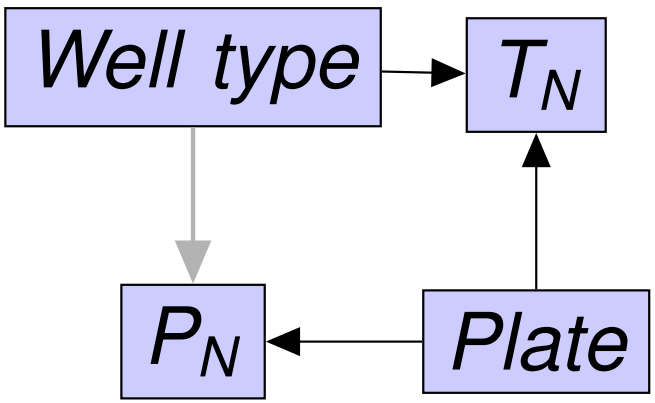
Graphical representation of effects



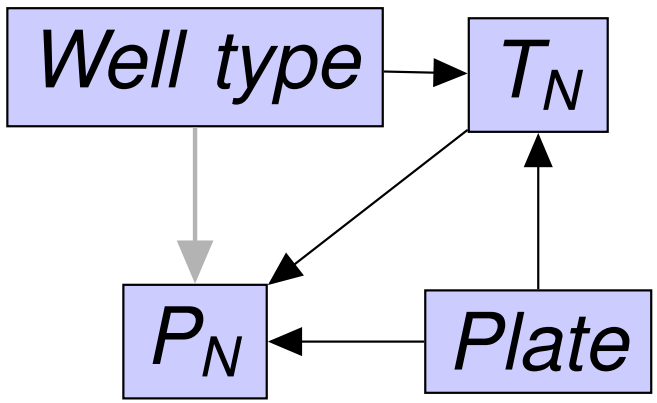
Graphical representation of effects



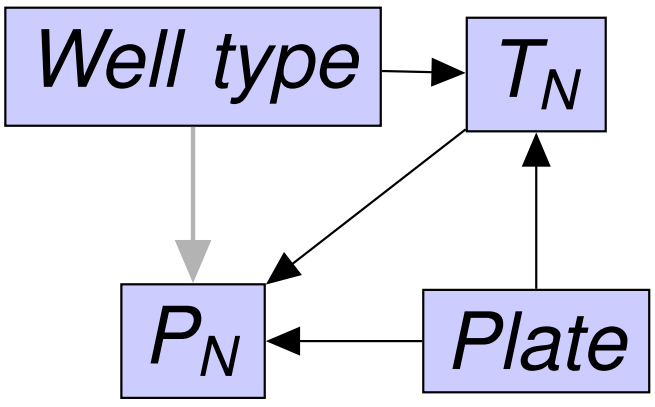
Graphical representation of effects



Graphical representation of effects



Graphical representation of effects

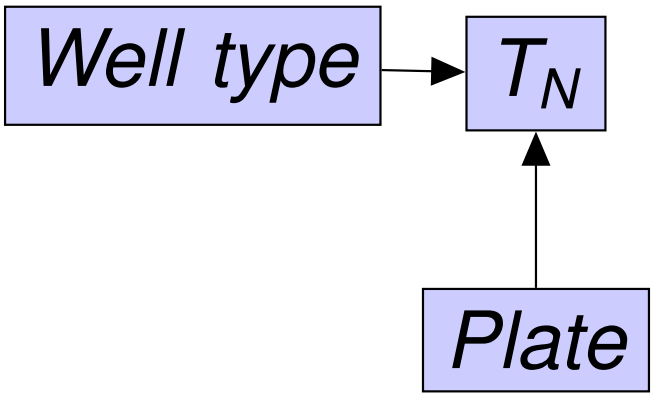



```
library(dagitty)

g1 <- dagitty( "dag {
    PN <- Well_type -> TN
    PN <- TN
    PN <- Plate -> TN
}")
```

```
adjustmentSets( g1, "Well_type",
    "PN", effect="direct" )
 { Plate, TN }

adjustmentSets( g1, "TN",
    "PN", effect="direct" )
 { Plate, Well_type }
```

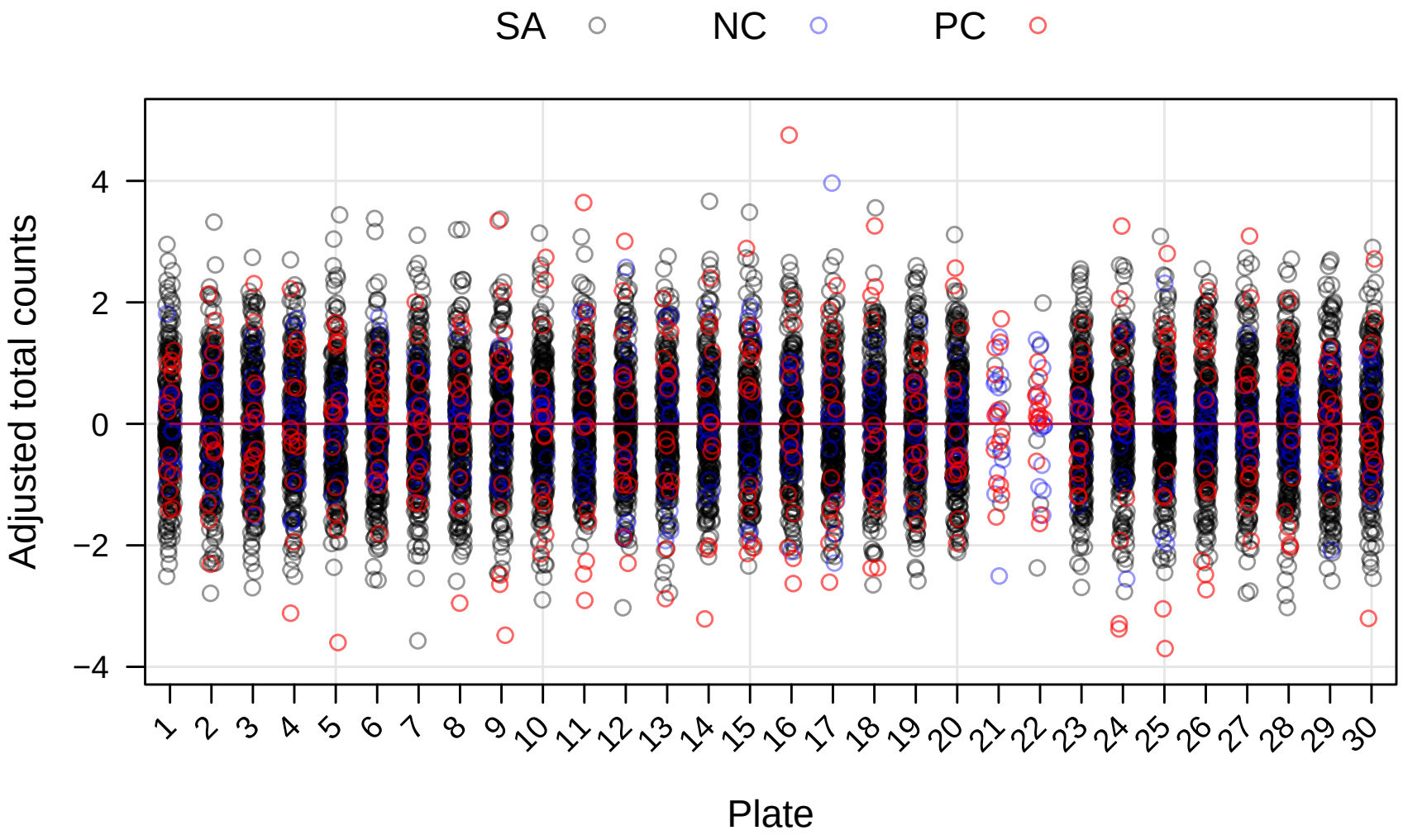
Fitting a model (2-Step approach)

Step 1: Adjust total cell count for plate and condition effects.



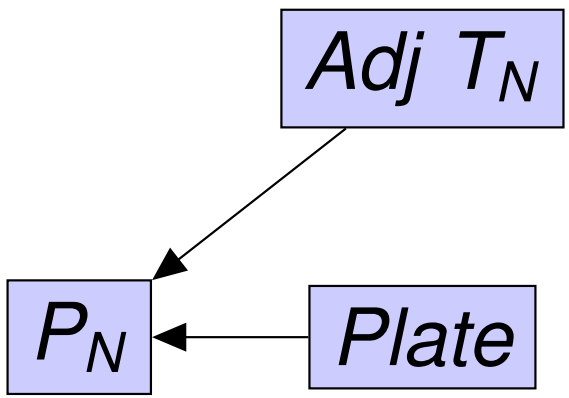

```
library(nlme)

gls(sqrt(TN) ~ Plate + Well_type +
                Plate:Well_type,
    weights=varIdent(~1 | Plate))
```

Adjusted total cell count



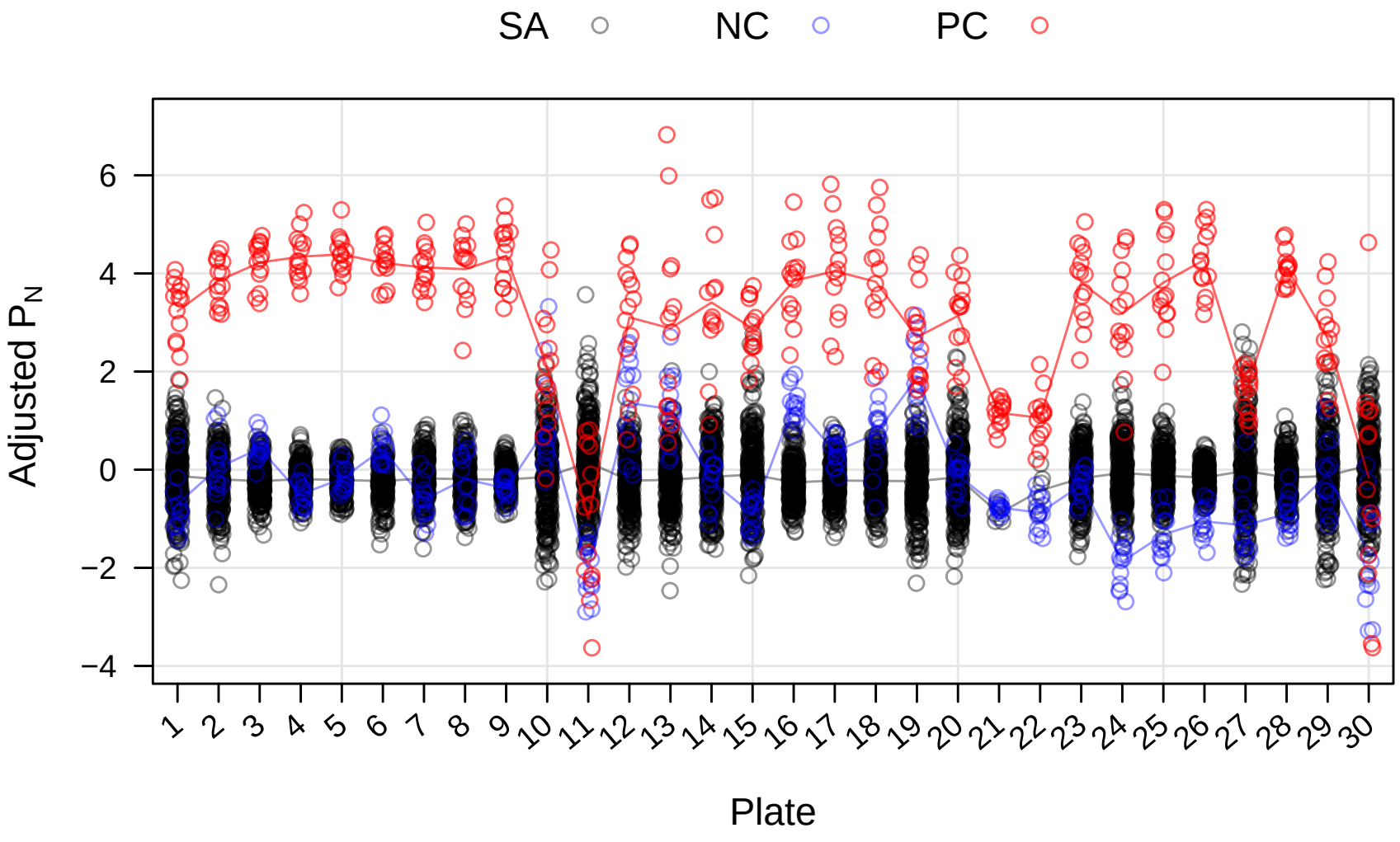
Fitting a model (2-Step approach)

Step 2: Adjust positive cell count for plate effects and (adjusted) total cell count.

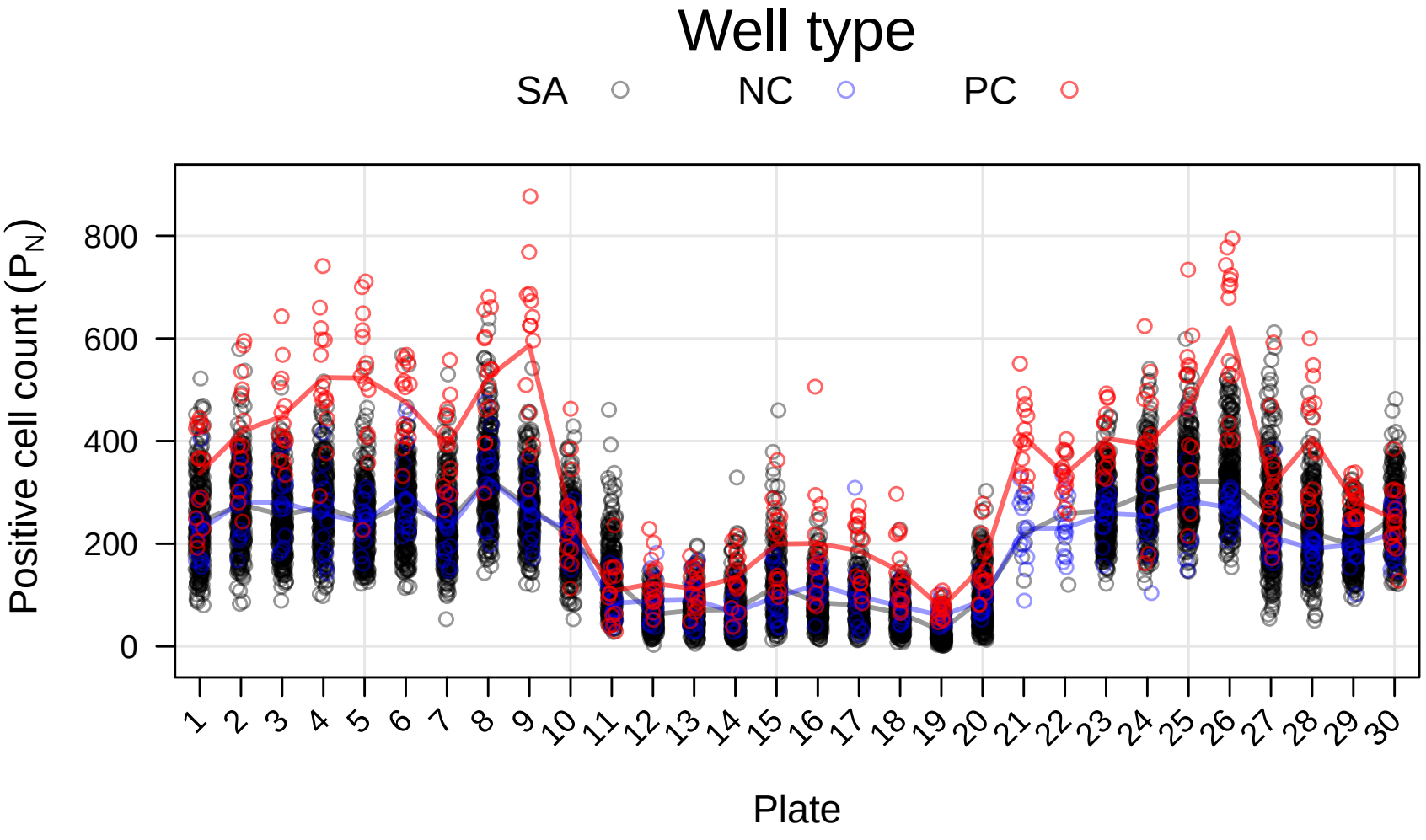



```
gls(sqrt(PN) ~ Plate + adj_TN,
    weights=varIdent(~1 | Plate)
```

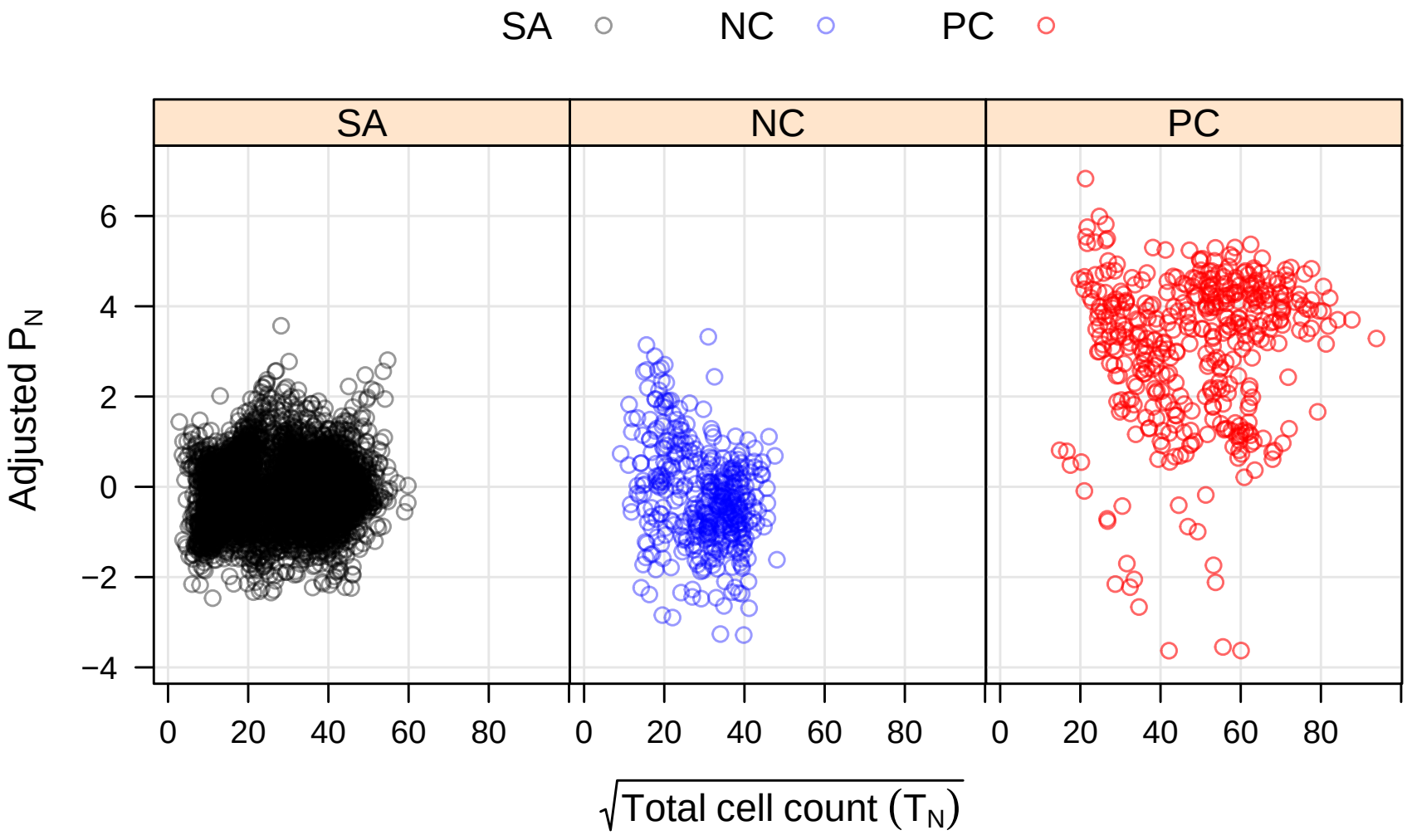
Adjusted positive cell count: plate effects removed



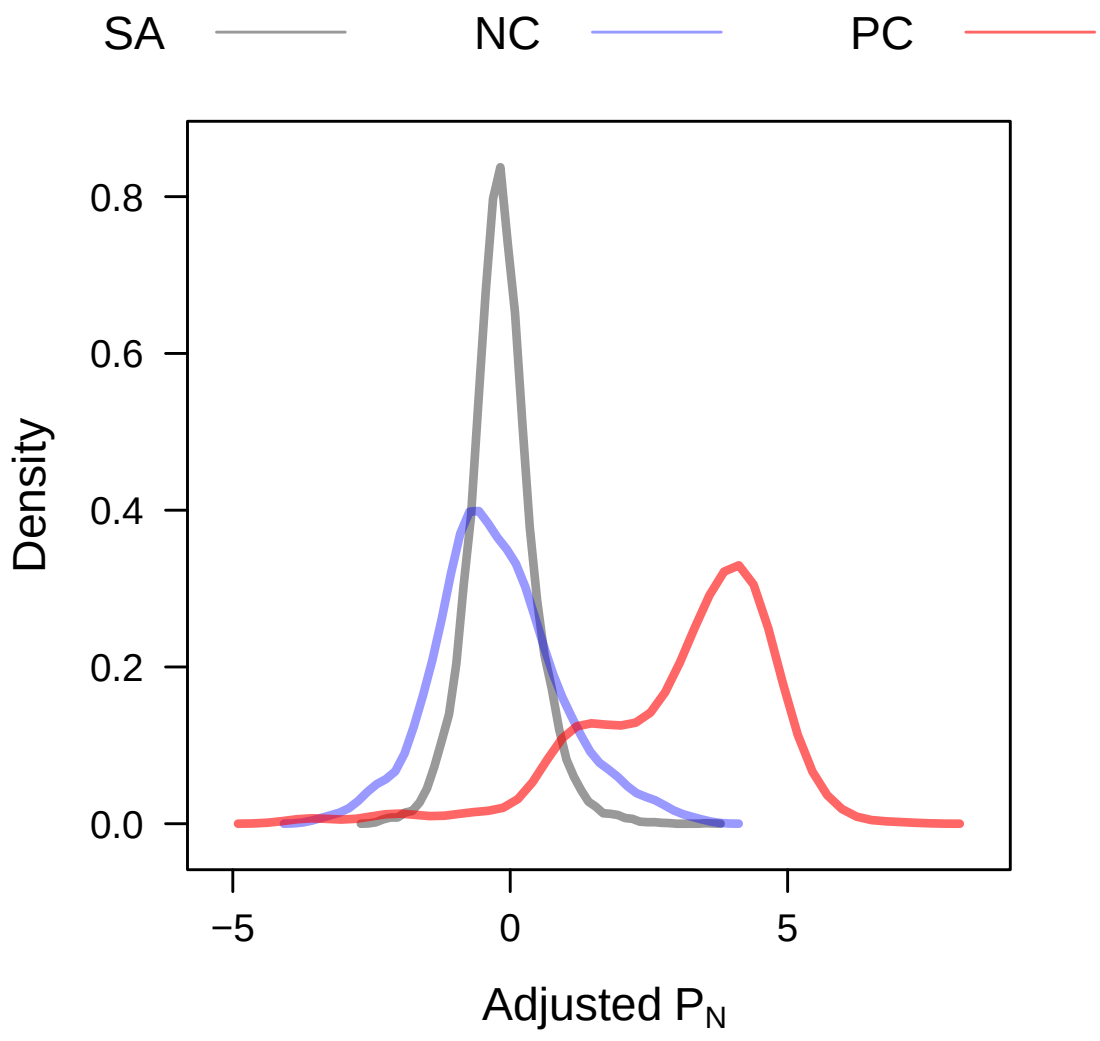
A look back at the unadjusted values



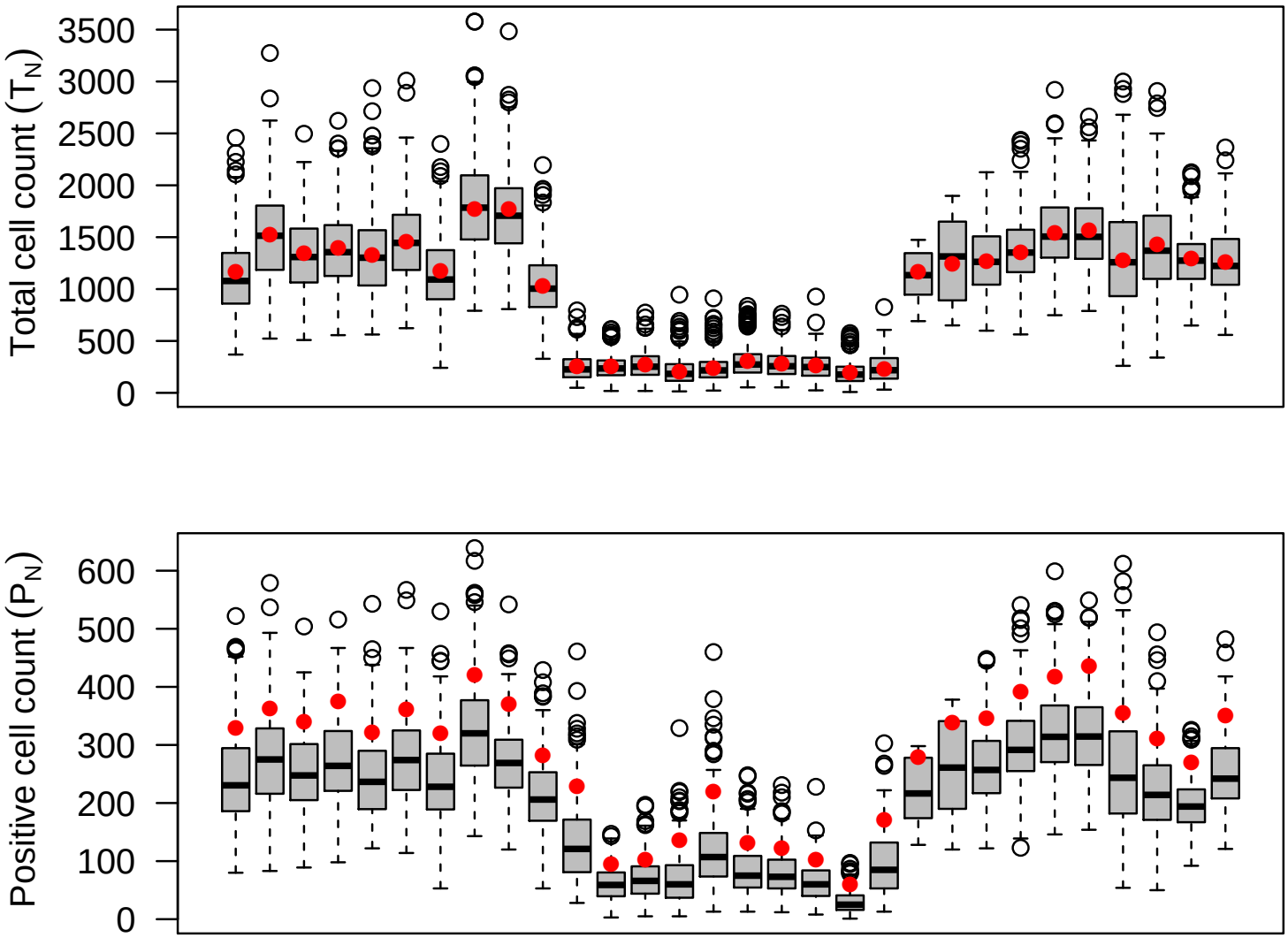
Dependence on total cell count removed



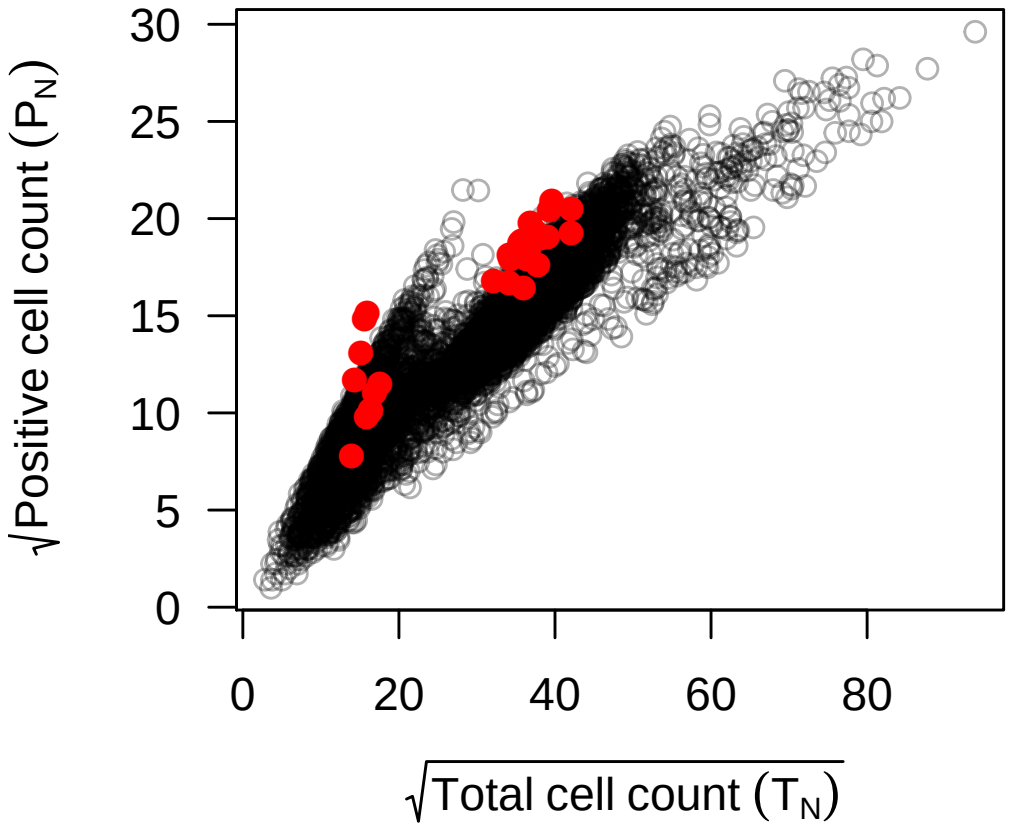
Positive controls > negative controls



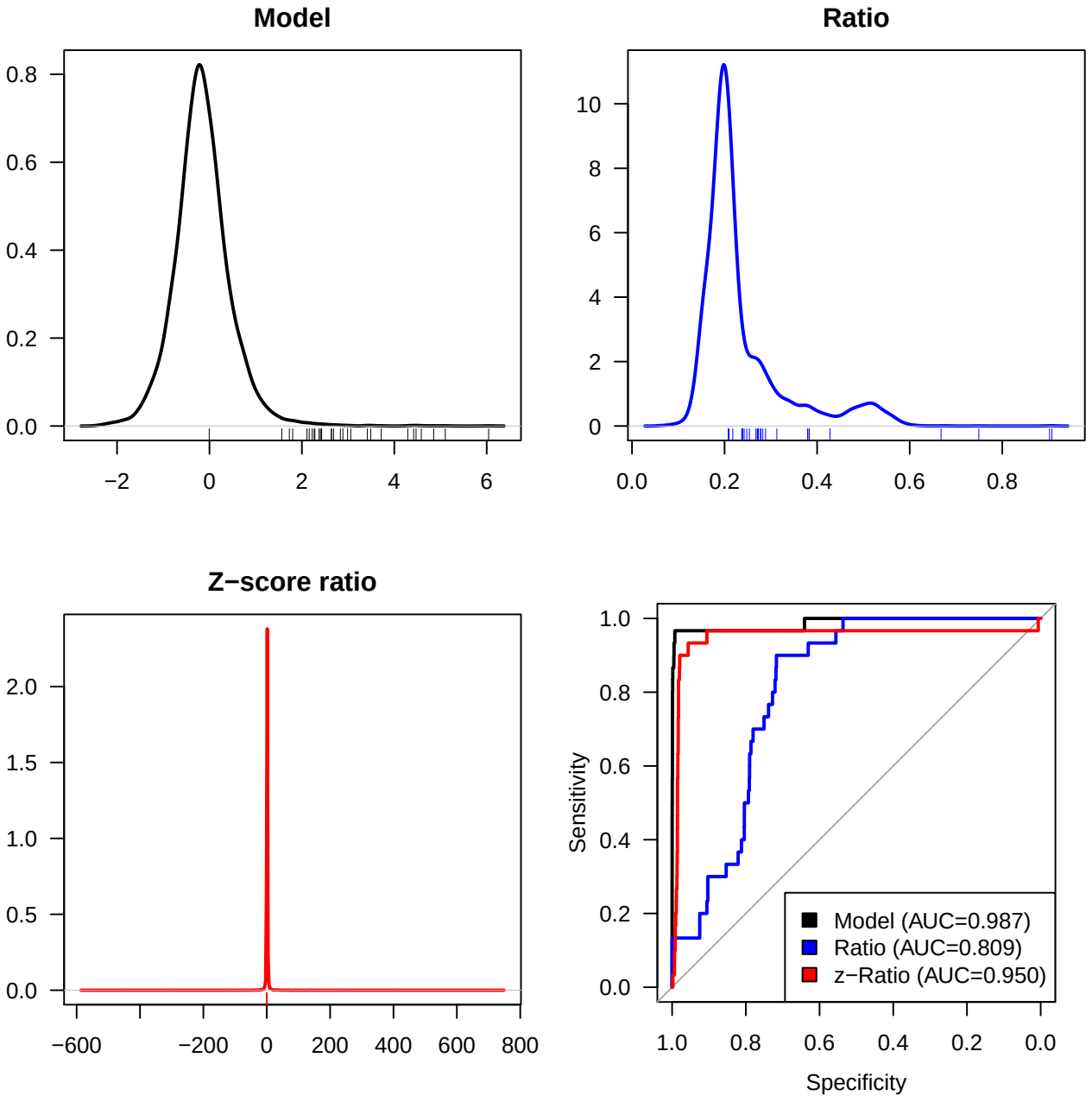
Spike-in controls: Mean T_N, 1.5 SD of mean P_N



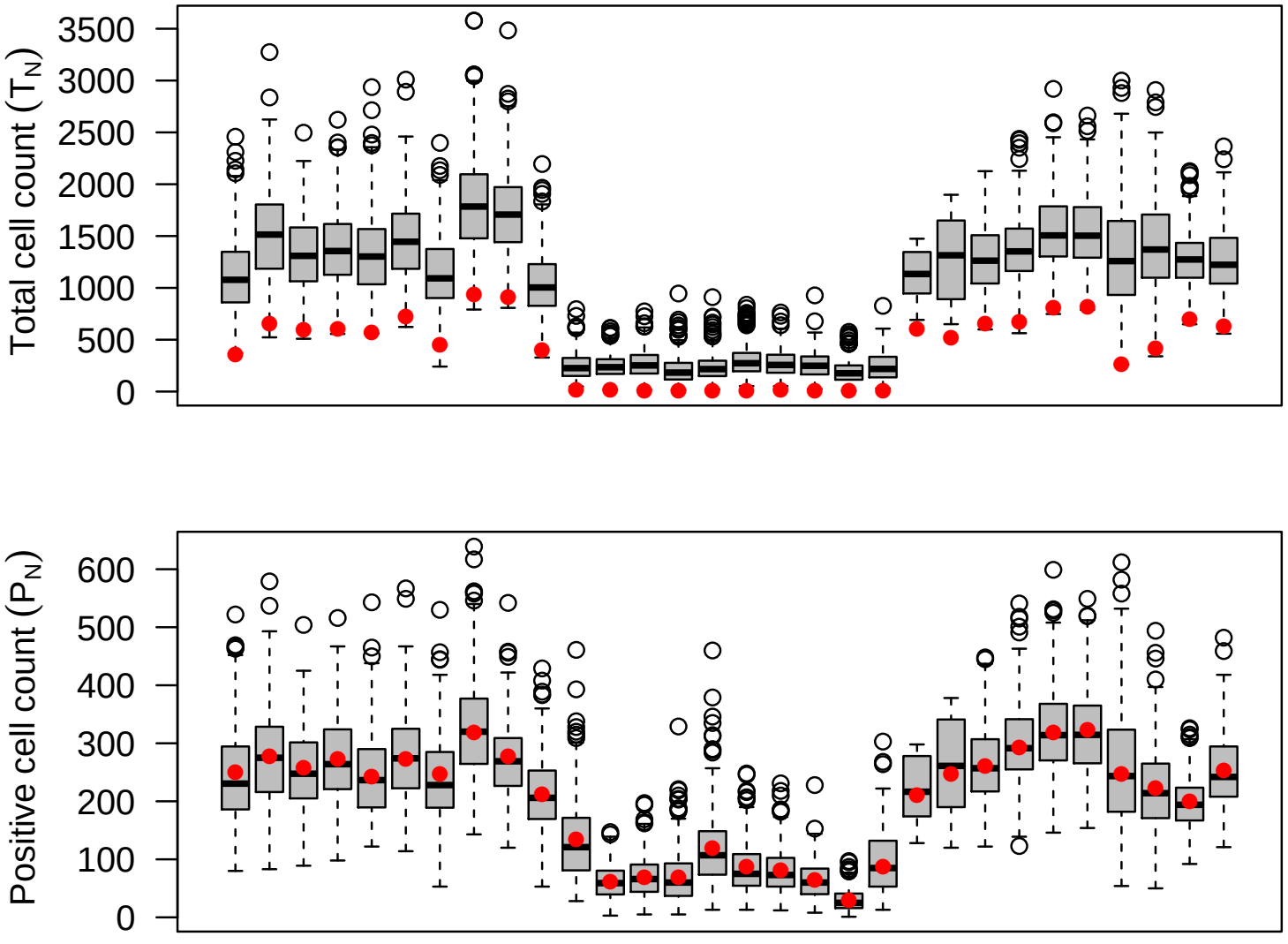
Spike-in controls: Mean T_N, 1.5 SD of mean P_N



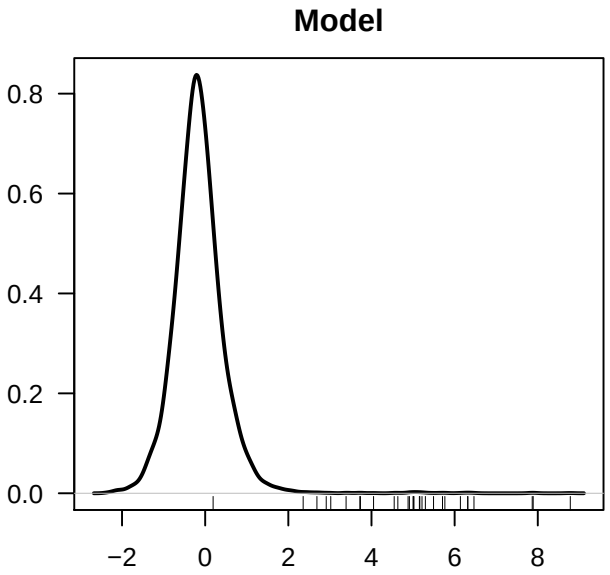
Spike-in controls

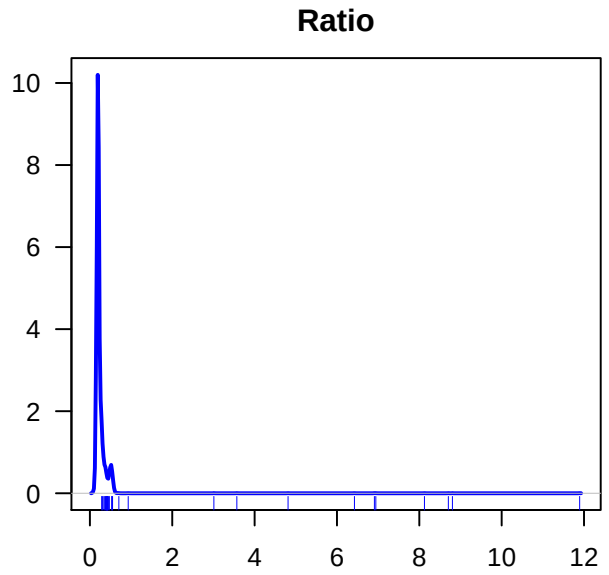


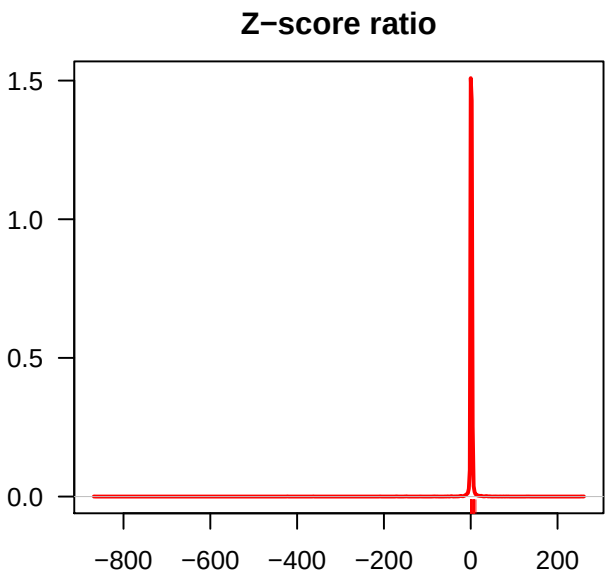
Spike-in controls: -2 SD of mean T_N, mean P_N

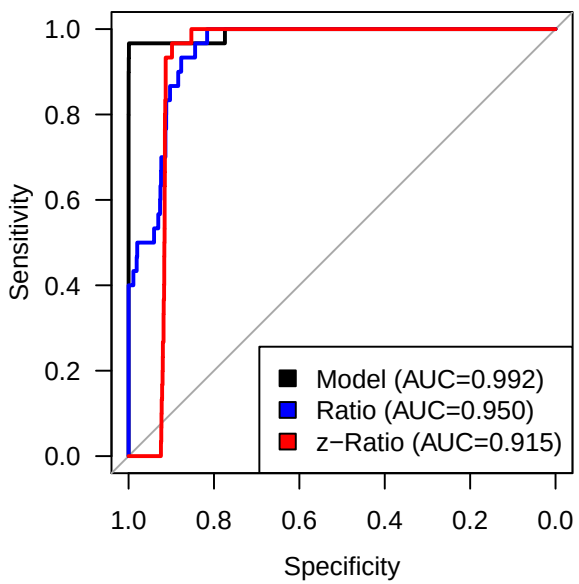


Spike-in controls









Assumptions, options, and extensions

- Count data can be suitably modelled as Gaussian $\rightarrow$ otherwise can use Poisson or negative binomial.
- Linear relationship between T_N and P_N (and constant across plates).
- Variances suitably modelled $\rightarrow$ what about separate variances for each well type within a plate?
- Would a hierarchical model perform better (e.g. treat plates random)?
- Would a one-step model perform better?
- How to incorporate spatial artefacts in the model?

Conclusions

- Removing artefacts and dependency on total cell count by fitting a model performs better than standard methods (ratio adjustments + normalising).
- But the key performance metric is if a preprocessing method improves hit calling.
- Statisticians should be involved in data preprocessing, not just the down-stream analysis.