Non-detects in qPCR data: methods to model and impute non-detects in the results of qPCR experiments (nondetects)

Matthew N. McCall

April 16, 2015

Contents

1 Background on non-detects in qPCR data **1**

2 EM algorithm **2**

3 Example **2**

4 Additional examples **6**

5 Session Info **7**

1 Background on non-detects in qPCR data

Quantitative real-time PCR (qPCR) measures gene expression for a subset of genes through repeated cycles of sequence-specific DNA amplification and expression measurements. During the exponential amplification phase, each cycle results in an approximate doubling of the quanitity of each target transcript. The threshold cycle (Ct) – the cycle at which the target gene's expression first exceeds a predetermined threshold – is used to quantify the expression of each target gene. These Ct values typically represent the raw data from a qPCR experiment.

One challenge of qPCR data is the presence of *non-detects* – those reactions failing to attain the expression threshold. While most current software replaces these non-detects with the maximum possible Ct value (typically 40), recent work has shown that this introduces large biases in estimation of both absolute and differential expression. Here, we treat the non-detects as missing data, model the missing data mechanism, and use this model to impute Ct values for the non-detects.

2 EM algorithm

We propose the following model of observed expression for gene i, sample-type j, and replicate k, Y_{ijk}:

$$Y_{ijk} = \begin{cases} \theta_{ij} + \delta_k + \varepsilon_{ijk} & \text{if } Z_{ijk} = 1 \\ \text{non-detect} & \text{if } Z_{ijk} = 0 \end{cases}$$

where δ_k represents a global shift in expression across samples and,

$$Pr(Z_{ijk} = 1) = \begin{cases} g(Y_{ijk}) & \text{if } Y_{ijk} < 40 \\ 0 & \text{otherwise} \end{cases}$$

Here, $g(Y_{ijk})$ can be estimated via the following logistic regression:

$$logit(Pr(Z_{ijk} = 1)) = \beta_0 + \beta_1 \hat{\theta}_{ij}$$

where $\hat{\theta}_{ij}$ is an estimate of the average expression for gene i and sample-type j.

3 Example

Data from Sampson *et al.* Oncogene 2013

Two cell types – young adult mouse colon (YAMC) cells and mutant-p53/activated-Ras transformed YAMC cells – in combination with three treatments – untreated, sodium butyrate, or valproic acid. Four replicates were performed for each cell-type/treatment combination [3].

Load the data

```
> library(HTqPCR)
> library(nondetects)
> data(oncogene2013)
```

Examine residuals when non-detects are replaced by 40

Normalize to Becn1:

```
> normCt <- normalizeCtData(oncogene2013, norm = "deltaCt",
+                           deltaCt.genes = "Becn1")
```

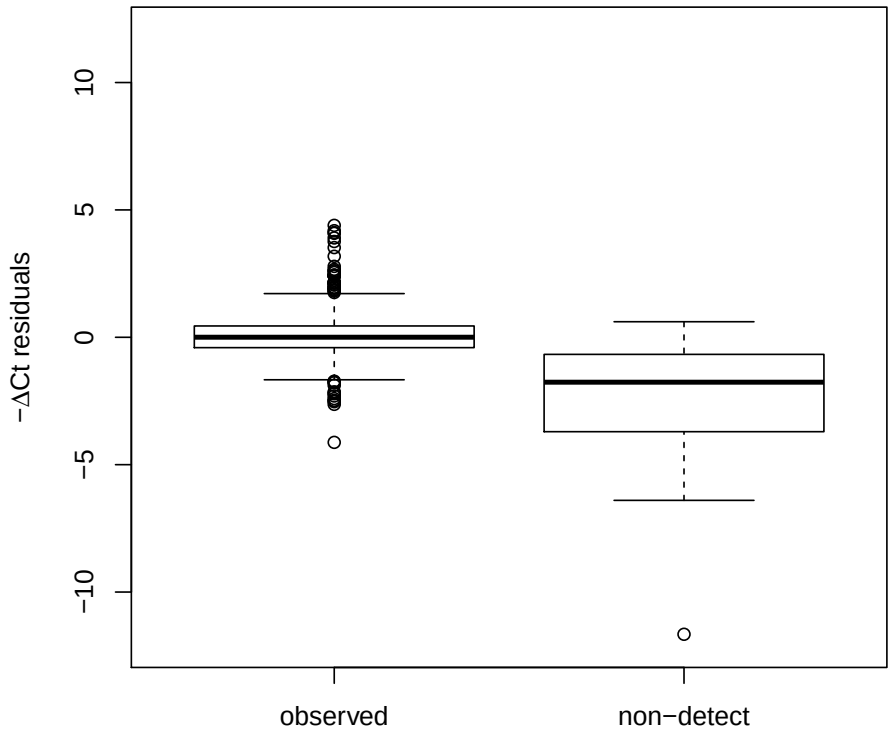
Calculate residuals for each set of replicates:

```
> conds <- paste(pData(normCt)$sampleType,pData(normCt)$treatment,sep=":")
> resids <- matrix(nrow=nrow(normCt), ncol=ncol(normCt))
> for(i in 1:nrow(normCt)){
+   for(j in 1:ncol(normCt)){
+     ind <- which(conds==conds[j])
+     resids[i,j] <- exprs(normCt)[i,j]-mean(exprs(normCt)[i,ind])
+   }
+ }
```

Create boxplots of residuals stratified by the presence of a non-detect:

```
> iND <- which(featureCategory(normCt)=="Undetermined", arr.ind=TRUE)
> iD <- which(featureCategory(normCt)!="Undetermined", arr.ind=TRUE)
> boxes <- list("observed"=-resids[iD], "non-detect"=-resids[iND])

> boxplot(boxes, main="",ylim=c(-12,12),
+         ylab=expression(paste("-",Delta,"Ct residuals",sep="")))
```



Impute non-detects

```
> oncogene2013 <- qpcrImpute(oncogene2013,
+                            groupVars=c("sampleType","treatment"))
```

Examine residuals when non-detects are replaced by imputed values

Normalize to Becn1:

```
> normCt <- normalizeCtData(oncogene2013, norm = "deltaCt",
+                           deltaCt.genes = "Becn1")
```

Remove the normalization gene:

```
> normCt <- normCt[-which(featureNames(normCt)=="Becn1"),]
```

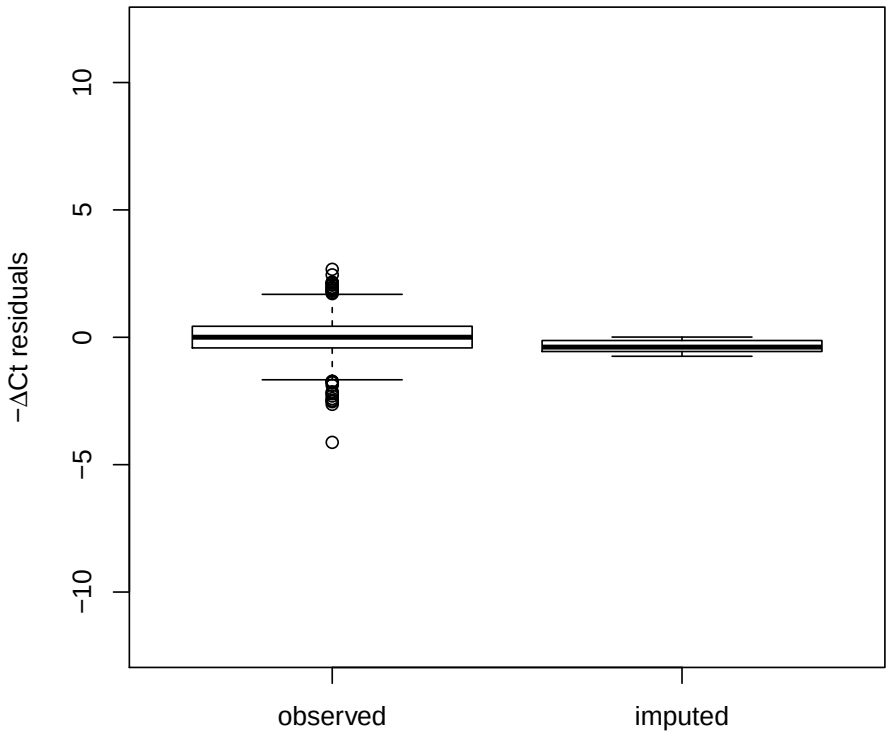
Calculate residuals for each set of replicates:

```
> conds <- paste(pData(normCt)$sampleType,
+                pData(normCt)$treatment,sep=":")
> resids <- matrix(nrow=nrow(normCt), ncol=ncol(normCt))
> for(i in 1:nrow(normCt)){
+   for(j in 1:ncol(normCt)){
+     ind <- which(conds==conds[j])
+     resids[i,j] <- exprs(normCt)[i,j]-mean(exprs(normCt)[i,ind])
+   }
+ }
```

Create boxplots of residuals stratified by the presence of a non-detect:

```
> iI <- which(featureCategory(normCt)=="Imputed", arr.ind=TRUE)
> iD <- which(featureCategory(normCt)!="Imputed", arr.ind=TRUE)
> boxes <- list("observed"=-resids[iD], "imputed"=-resids[iI])

> boxplot(boxes, main="",ylim=c(-12,12),
+         ylab=expression(paste("-",Delta,"Ct residuals",sep="")))
```



4 Additional examples

Two additional example data sets are used in the paper and included in the package. These are each briefly described below.

Data from Almudevar *et al.* SAGMB 2011

Cells transformed to malignancy by mutant p53 and activated Ras are perturbed with the aim of restoring gene expression to levels found in non-transformed parental cells via retrovirus-mediated re-expression of corresponding cDNAs or shRNA-dependent stable knock-down. The data contain 4-6 replicates for each perturbation, and each perturbation has a corresponding control sample in which only the vector has been added [1].

```
> library(nondetects)
> data(sagmb2011)
```

Data from McMurray *et al.* Nature 2008

A study of the effect of p53 and/or Ras mutations on gene expression. The third dataset is a comparison between four cell types – YAMC cells, mutant-p53 YAMC cells, activated-Ras YAMC cells, and p53/Ras double mutant YAMC cells. Three replicates were performed for the untransformed YAMC cells, and four replicates were performed for each of the other cell types [2].

```
> library(nondetects)
> data(nature2008)
```

Funding

This work was supported by National Institutes of Health [grant numbers CA009363, CA138249, HG006853]; and an Edelman-Gardner Foundation Award.

5 Session Info

```
> sessionInfo()
```

```
R version 3.2.0 (2015-04-16)
Platform: x86_64-unknown-linux-gnu (64-bit)
Running under: Ubuntu 14.04.2 LTS

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] parallel  stats     graphics  grDevices utils     datasets  methods
```

```
[8] base

other attached packages:
[1] nondetects_1.4.0    HTqPCR_1.22.0       limma_3.24.0
[4] RColorBrewer_1.1-2  Biobase_2.28.0      BiocGenerics_0.14.0

loaded via a namespace (and not attached):
 [1] gtools_3.4.2          bitops_1.0-6          affy_1.46.0
 [4] stats4_3.2.0          KernSmooth_2.23-14    BiocInstaller_1.18.1
 [7] gplots_2.16.0         zlibbioc_1.14.0       gdata_2.13.3
[10] affyio_1.36.0         preprocessCore_1.30.0 tools_3.2.0
[13] caTools_1.17.1
```

References

[1] A. Almudevar, M. N. McCall, H. McMurray, and H. Land. Fitting Boolean networks from steady state perturbation data. *Statistical applications in genetics and molecular biology*, 10(1):47, 2011.

[2] H. R. McMurray, E. R. Sampson, G. Compitello, C. Kinsey, L. Newman, B. Smith, S.-R. Chen, L. Klebanov, P. Salzman, A. Yakovlev, and H. Land. Synergistic response to oncogenic mutations defines gene class critical to cancer phenotype. *Nature*, 453(7198):1112–1116, 2008.

[3] E. Sampson, H. McMurray, D. Hassane, L. Newman, P. Salzman, C. Jordan, and H. Land. Gene signature critical to cancer phenotype as a paradigm for anticancer drug discovery. *Oncogene*, 32(33):3809–18, 2013.