Case Studies in Interoperability: From Generic Classes to Specific Functions

Matthew N. McCall

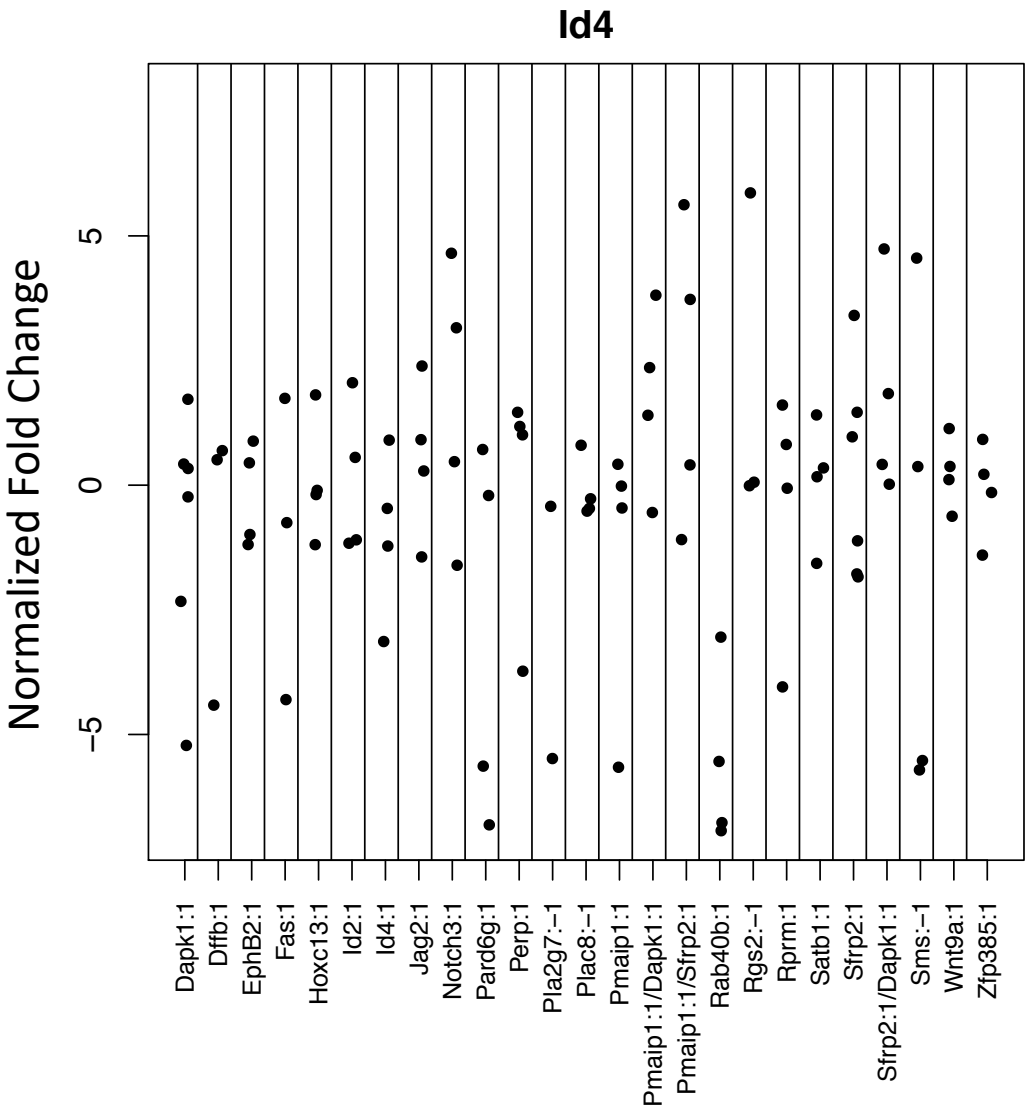
Department of Biostatistics & Computational Biology

University of Rochester Medical Center

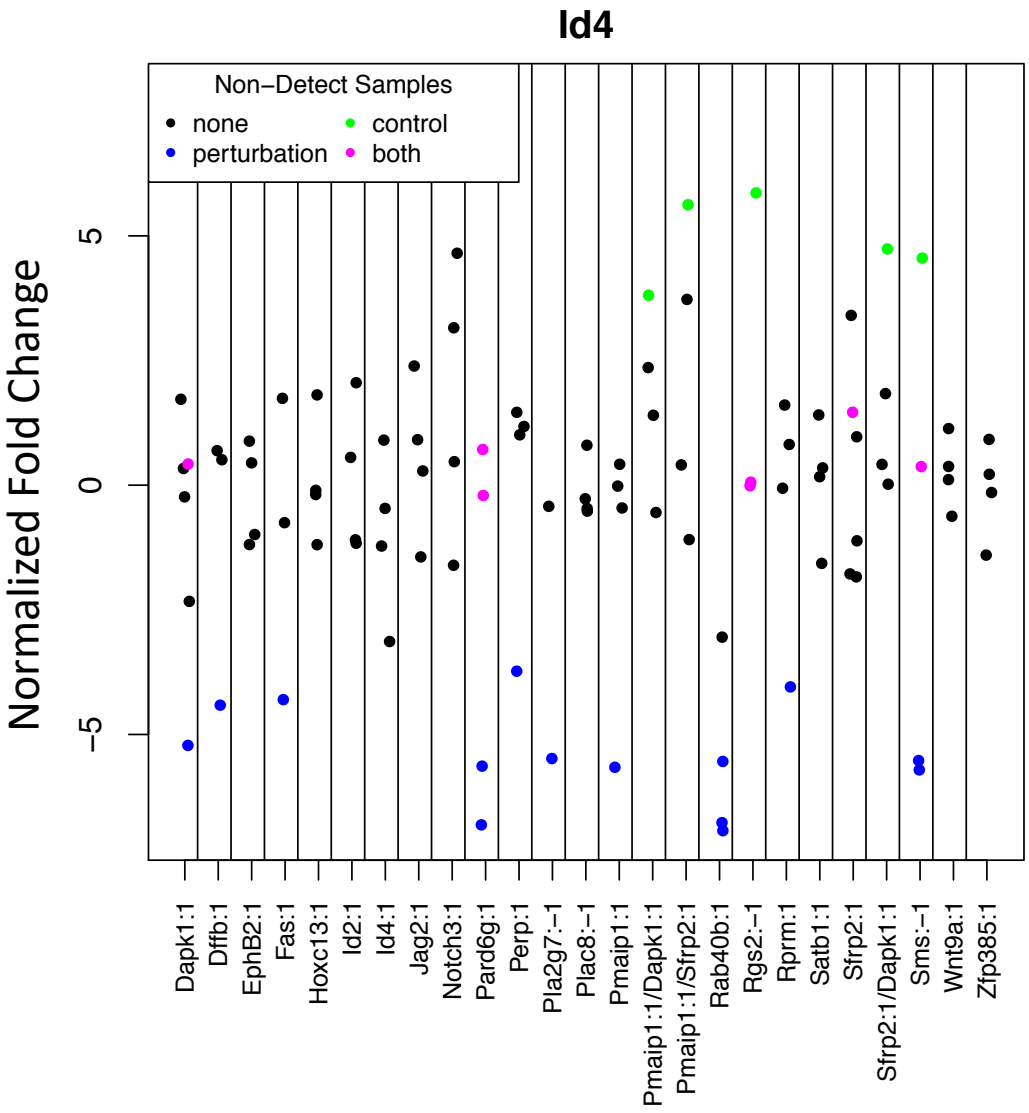
`mnmccall.com`

`@matthewnmccall`

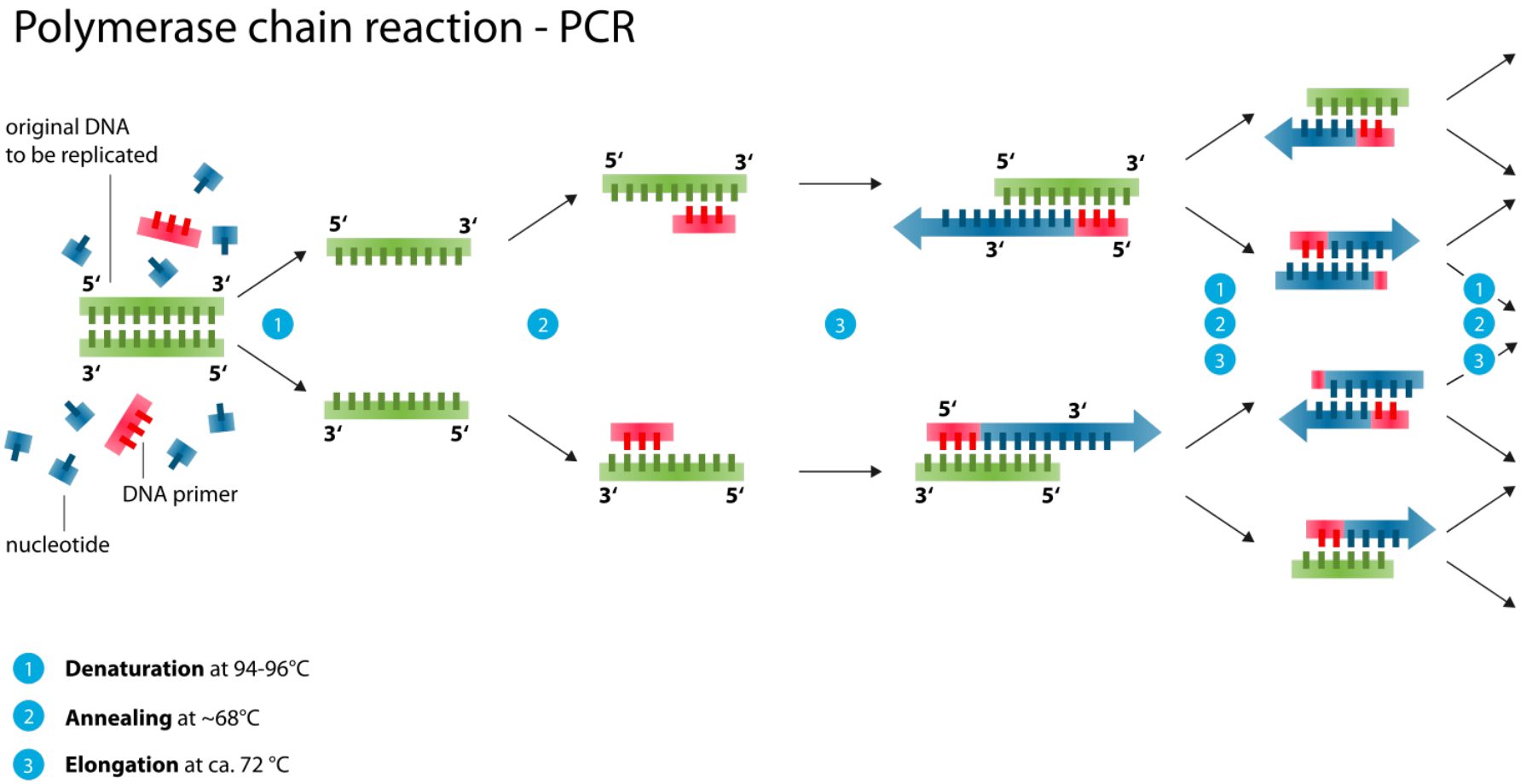
Origin Story: gene perturbations and cancer systems biology



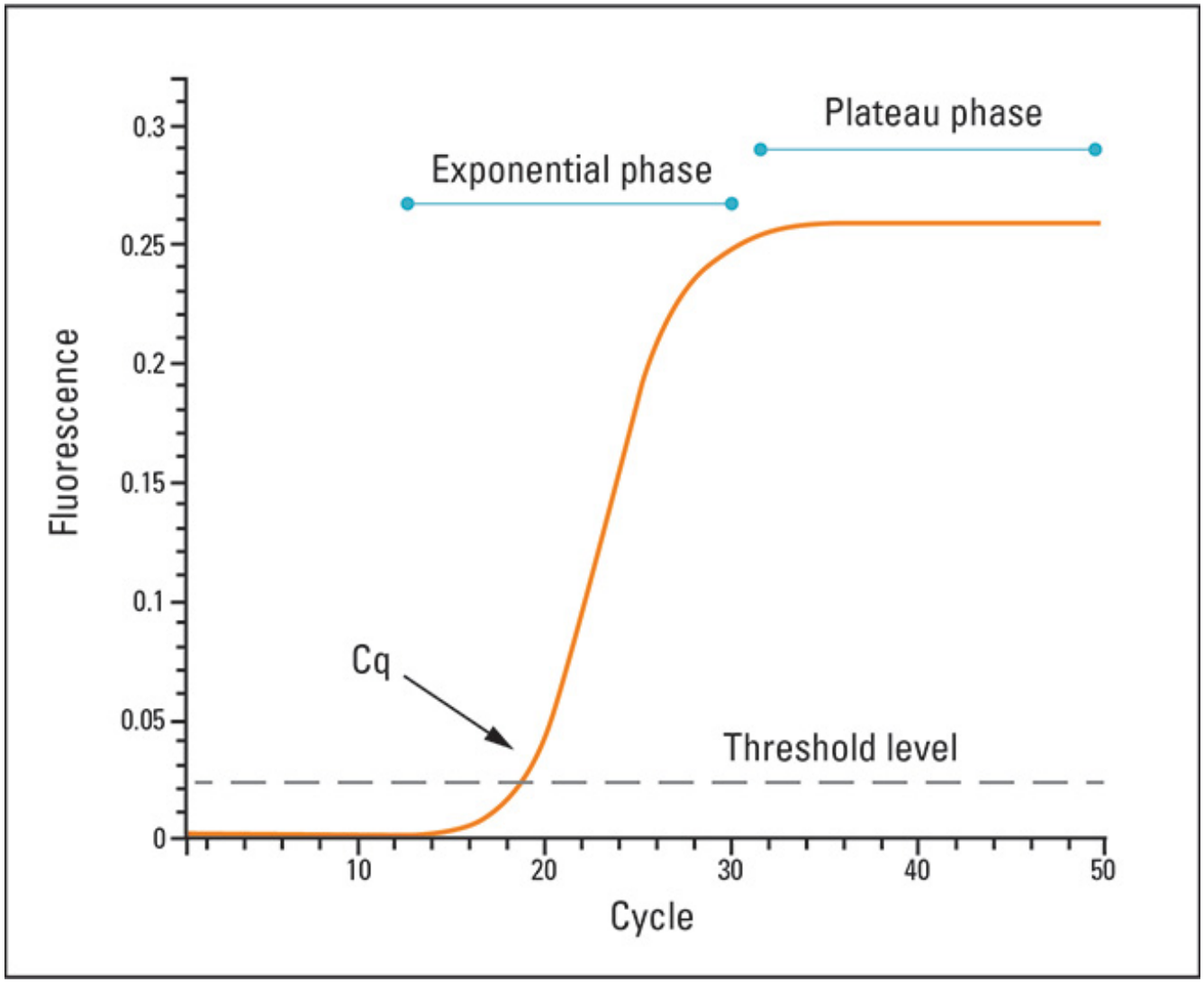
Outliers mostly due to non-detects



A quick intro to PCR

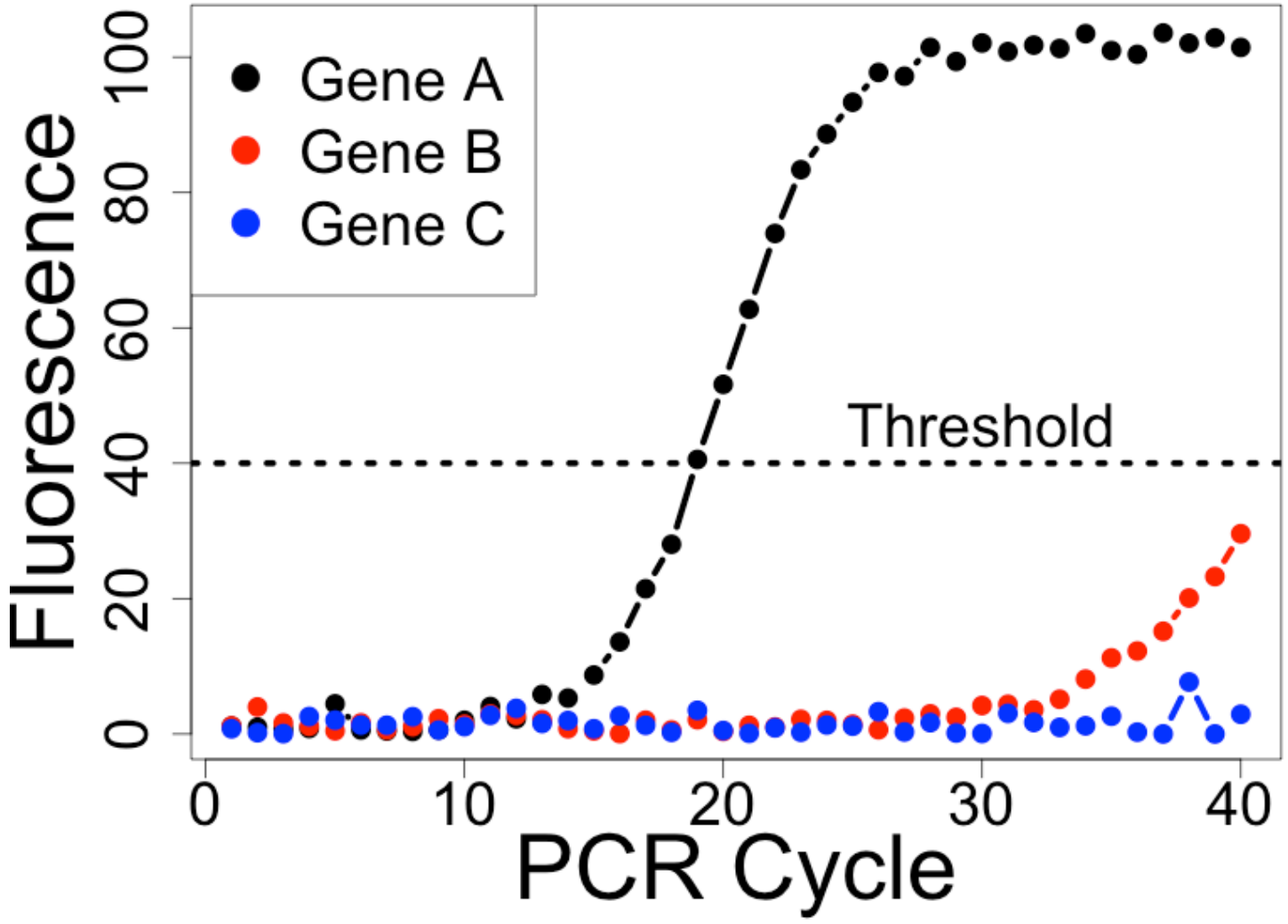


A quick intro to qPCR

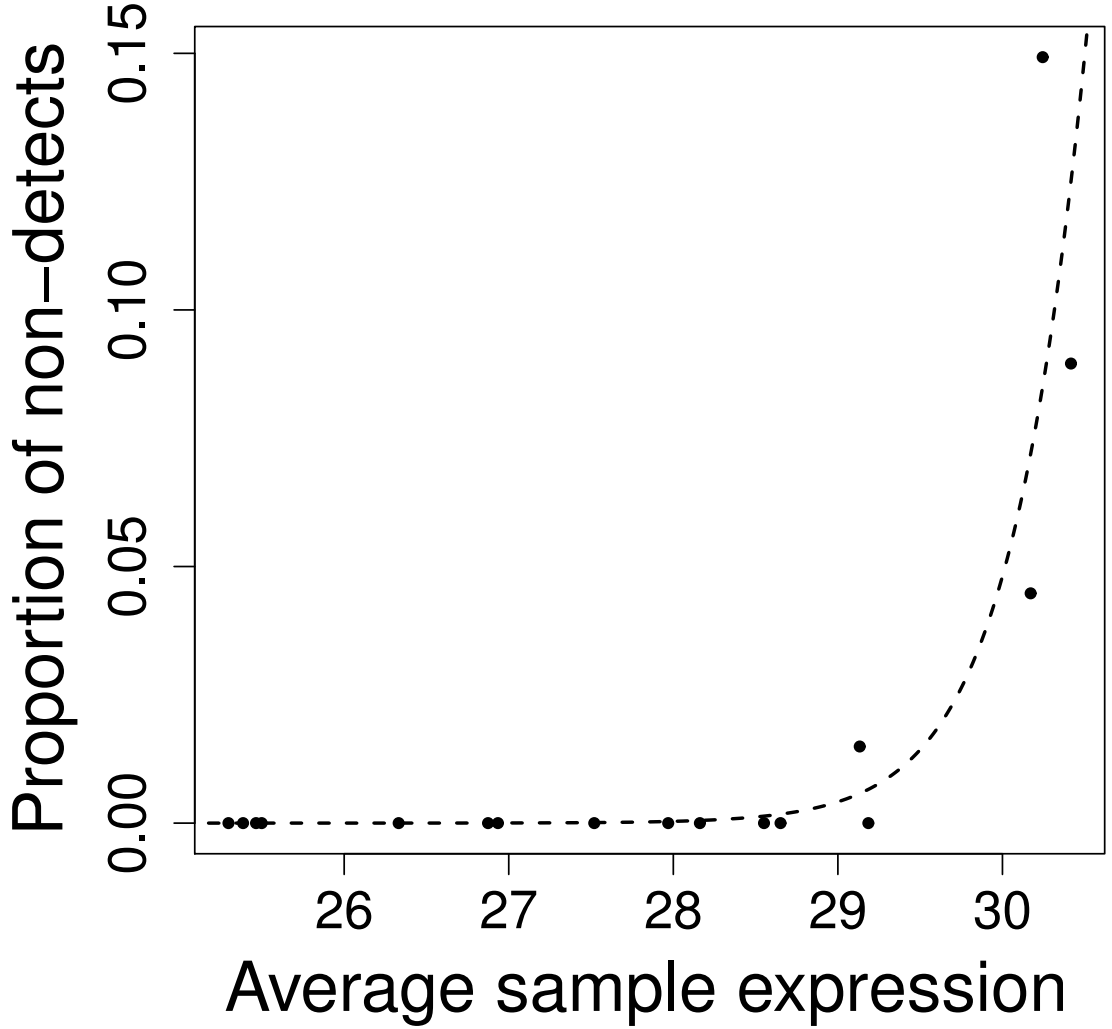


Thermo Scientific, Basic Principles of qPCR

Non-detects in qPCR



Non-detects do not occur randomly



i gene
j condition
k sample

Impute non-detects based on the following model:

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

$$\varepsilon_{ijk} \sim \mathbf{N}(0, \sigma^2)$$

where θ_{ij} is gene expression and δ_k represents an array effect.

i gene
j condition
k sample

Impute non-detects based on the following model:

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

$$\varepsilon_{ijk} \sim \mathbf{N}(0, \sigma^2)$$

where θ_{ij} is gene expression and δ_k represents an array effect.

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

i gene
j condition
k sample

Impute non-detects based on the following model:

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

$$\varepsilon_{ijk} \sim \mathbf{N}(0, \sigma^2)$$

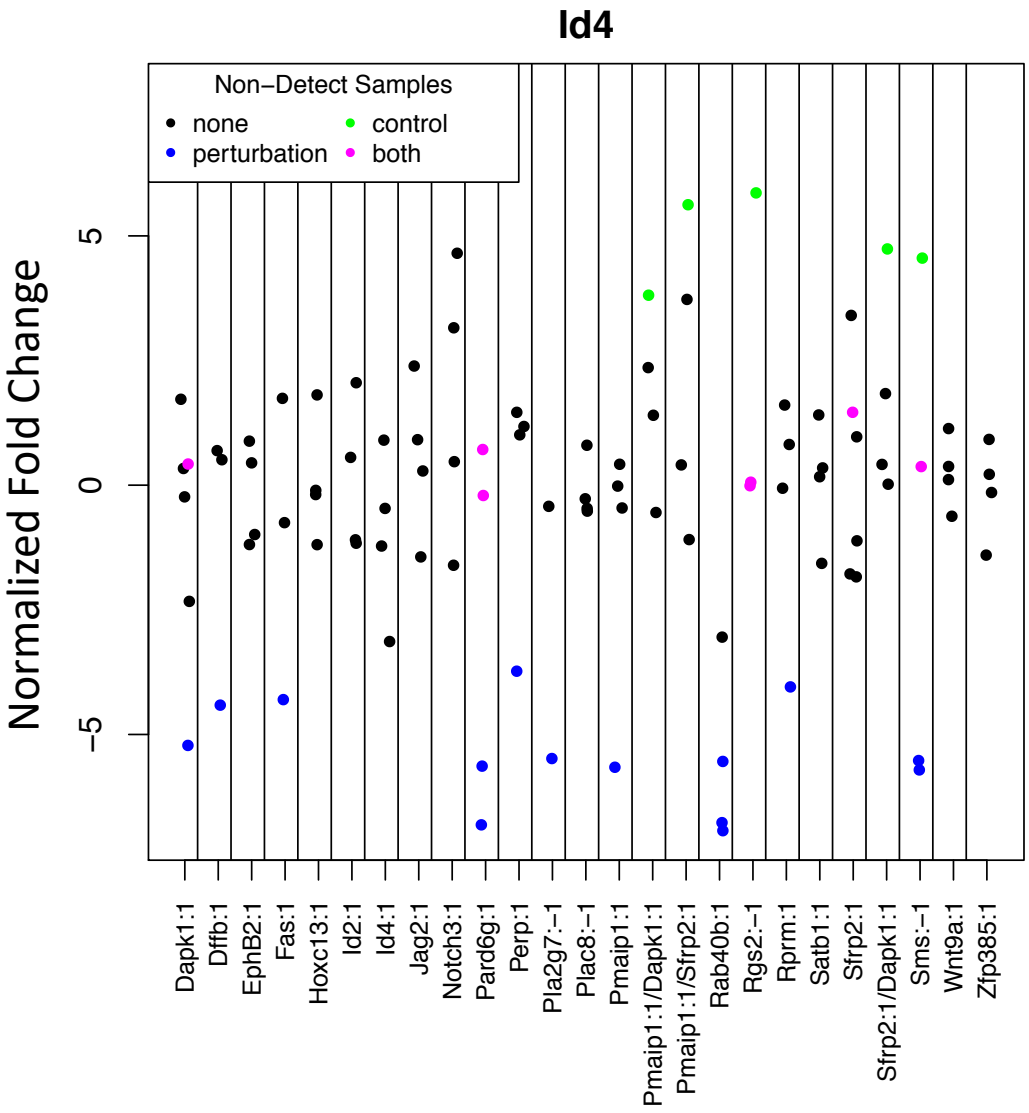
where θ_{ij} is gene expression and δ_k represents an array effect.

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

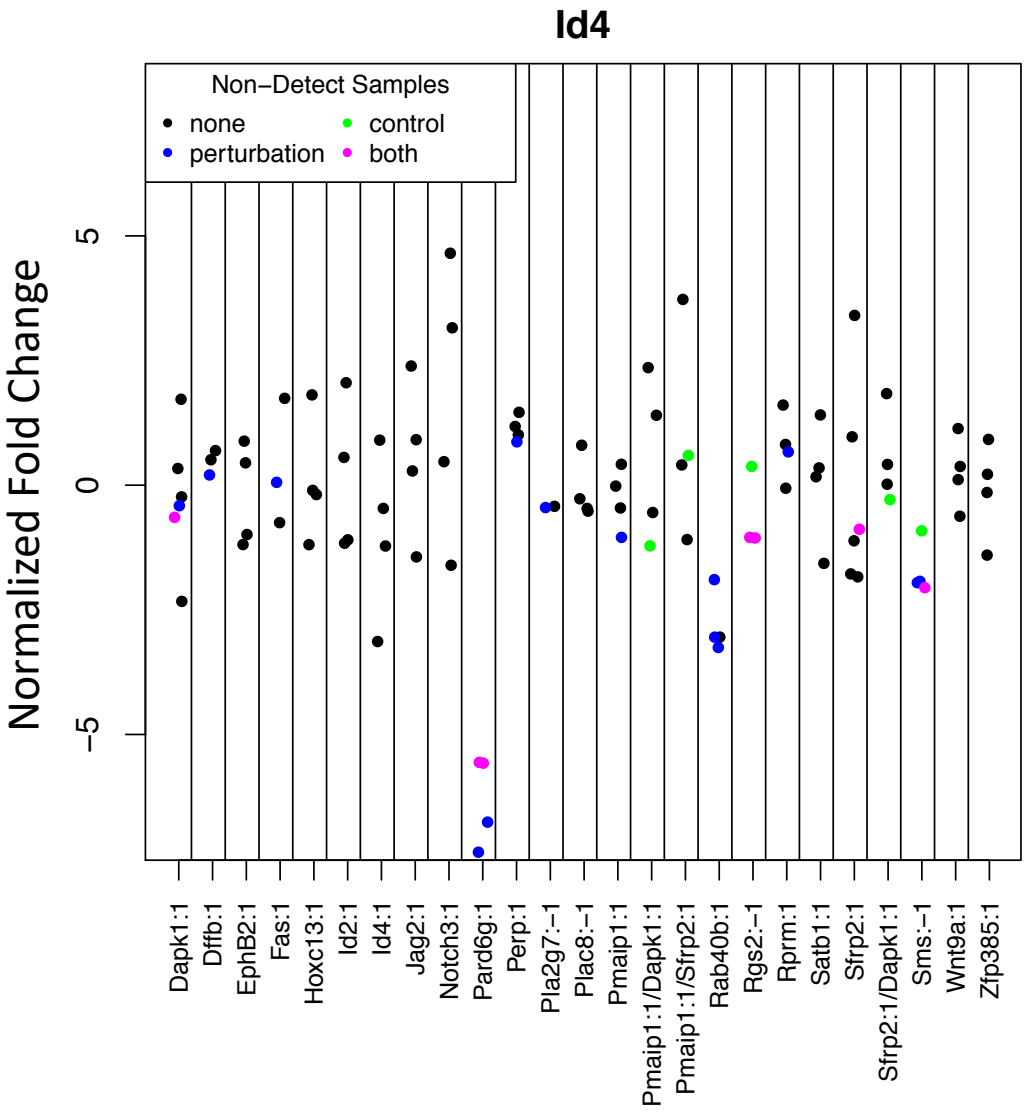
For non-detects:

$$\hat{Y}_{ijk} = \mathbf{E}\left\{Y_{ijk} \mid \text{non-detect}; \theta_{ij}, \delta_k, \sigma^2\right\}$$

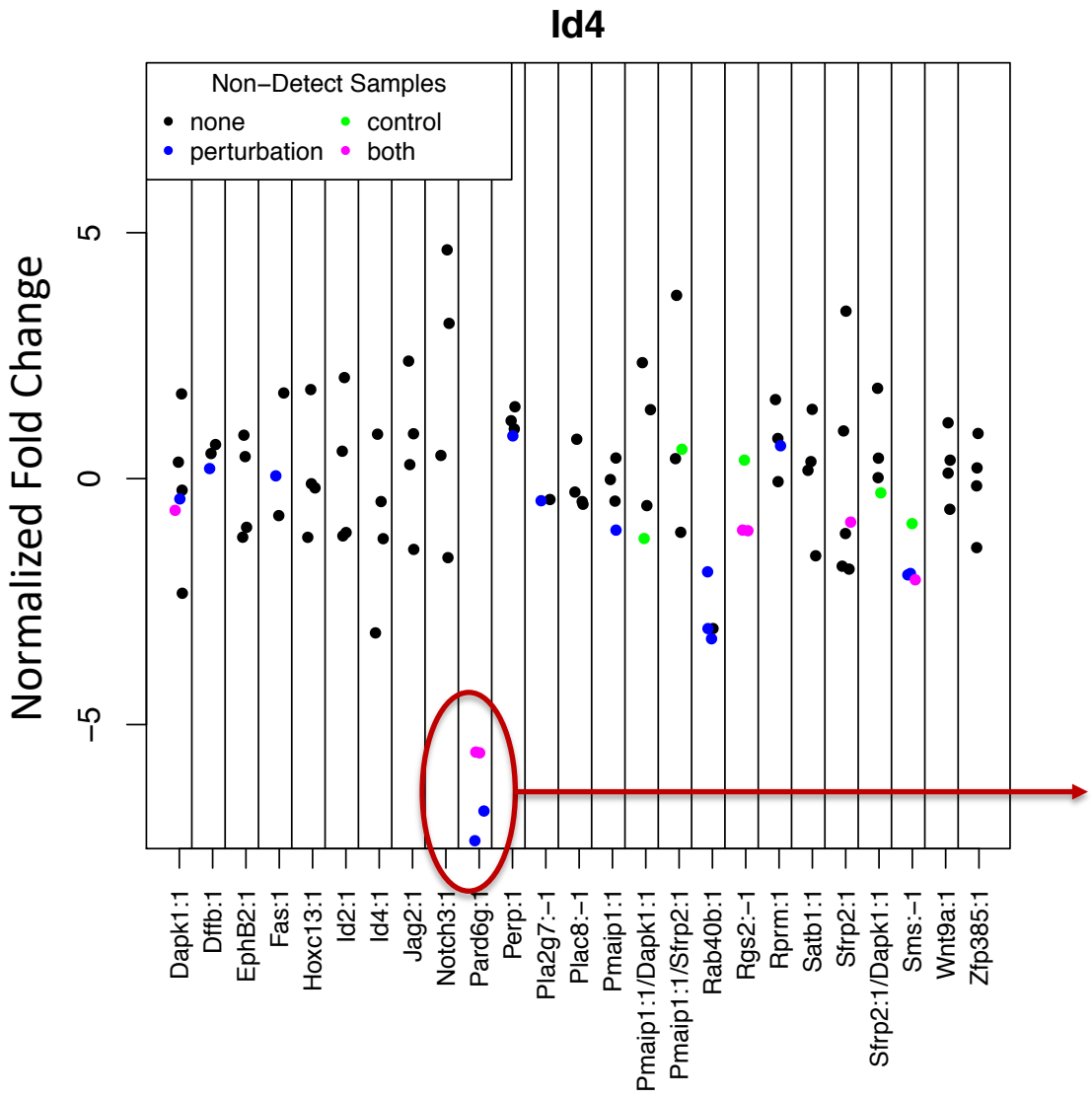
Remember this figure



Farewell to (most) outliers



Farewell to (most) outliers



Need something fancier when all replicates have a non-detect.

Further Reading

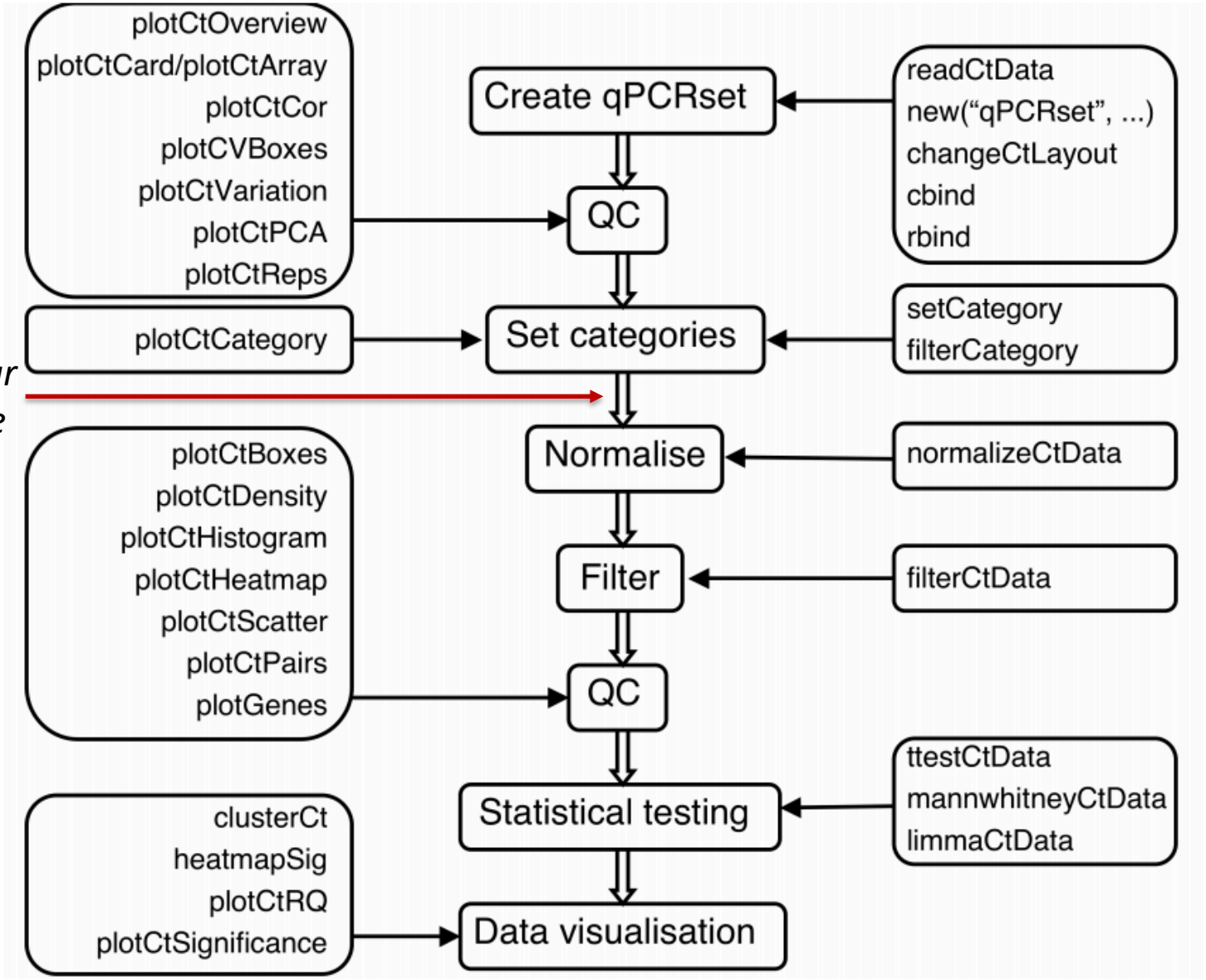
McCall, M. N., McMurray, H. R., Land, H., & Almudevar, A. (2014). On non-detects in qPCR data. *Bioinformatics*, *30*(16), 2310-2316.

Sherina, V., McMurray, H., Powers, W., Land, H., Love, T., & McCall, M. N. (2017). Statistical Approaches to Decreasing the Discrepancy of Non-detects in qPCR Data. *bioRxiv*, 231621.

We have a new method for one small part of an analysis pipeline.

Analysis of qPCR data

Want to insert our new method here



HTqPCR

platforms all | rank unknown | posts 2 / 0 / 0 / 0 | in Bioc 9.5 years
build warnings | updated before release

DOI: 10.18129/B9.bioc.HTqPCR

Automated analysis of high-throughput qPCR data

Bioconductor version: Release (3.9)

Analysis of Ct values from high throughput quantitative real-time PCR (qPCR) assays across multiple conditions or replicates. The input data can be from spatially-defined formats such ABI TaqMan Low Density Arrays or OpenArray; LightCycler from Roche Applied Science; the CFX plates from Bio-Rad Laboratories; conventional 96- or 384-well plates; or microfluidic devices such as the Dynamic Arrays from Fluidigm Corporation. HTqPCR handles data loading, quality assessment, normalization, visualization and parametric or non-parametric testing for statistical significance in Ct values between features (e.g. genes, microRNAs).

Author: Heidi Dvinge, Paul Bertone

Maintainer: Heidi Dvinge <hdvinge at fredhutch.org>

Citation (from within R, enter `citation("HTqPCR")`):

Dvinge H, Bertone P (2009). "HTqPCR: High - throughput analysis and visualization of quantitative real - time PCR data in R." *Bioinformatics*, **25(24)**, 3325.

nondetects

platforms all | rank unknown | posts 1 / 0 / 0 / 0 | in Bioc 5 years

build ok | updated before release

DOI: 10.18129/B9.bioc.nondetects

Non-detects in qPCR data

Bioconductor version: Release (3.9)

Methods to model and impute non-detects in the results of qPCR experiments.

Author: Matthew N. McCall <mccallm at gmail.com>, Valeriia Sherina <valery.sherina at gmail.com>

Maintainer: Valeriia Sherina <valery.sherina at gmail.com>

Citation (from within R, enter `citation("nondetects")`):

McCall MN, McMurray H, Land H, Almudevar A (2014). "On Non-detects in qPCR Data." *Bioinformatics*.

Details

biocViews	AssayDomain, GeneExpression, Preprocessing, Software, Technology, WorkflowStep, qPCR
Version	2.14.0
In Bioconductor since	BioC 2.14 (R-3.1) (5 years)
License	GPL-3
Depends	R (>= 3.2), Biobase(>= 2.22.0)
Imports	limma, mvtnorm, utils, methods, arm, HTqPCR(>= 1.16.0)
LinkingTo	
Suggests	knitr, rmarkdown, BiocStyle(>= 1.0.0), RUnit, BiocGenerics(>= 0.8.0)

```
> library(nondetects)
> data(sagmb2011)
> sagmb2011
An object of class "qPCRset"
Size:  67 features, 55 samples
Feature types:
Feature names:           Abat Abca1 Ank ...
Feature classes:
Feature categories:      OK, Undetermined
Sample names:            CK1.Vector CK2.Vector DD2.Vector ...
```

```
> library(nondetects)
> data(sagmb2011)
> sagmb2011
An object of class "qPCRset"
Size:  67 features, 55 samples
Feature types:
Feature names:           Abat Abca1 Ank ...
Feature classes:
Feature categories:      OK, Undetermined
Sample names:            CK1.Vector CK2.Vector DD2.Vector ...


> sagmb2011 <- qpcrImpute(sagmb2011, groupVars="sampleType")
~0 + nrep
<environment: 0x11ad6ef00>
[1] "1 / 100"
-5724.6243728202
[1] "2 / 100"
-5692.91657561198
[1] "3 / 100"
-5685.33706725388
[1] "4 / 100"
-5681.66921373554
[1] "5 / 100"
-5679.73392983577
[1] "6 / 100"
-5678.7036584038
[1] "7 / 100"
-5678.15423981277
[1] "Single"
```

```
> library(nondetects)
> data(sagmb2011)
> sagmb2011
An object of class "qPCRset"
Size:  67 features, 55 samples
Feature types:
Feature names:           Abat Abca1 Ank ...
Feature classes:
Feature categories:      OK, Undetermined
Sample names:            CK1.Vector CK2.Vector DD2.Vector ...


> sagmb2011 <- qpcrImpute(sagmb2011, groupVars="sampleType")


> sagmb2011
An object of class "qPCRset"
Size:  67 features, 55 samples
Feature types:
Feature names:           Abat Abca1 Ank ...
Feature classes:
Feature categories:      OK, Imputed
Sample names:            CK1.Vector CK2.Vector DD2.Vector ...
```

```
> library(nondetects)
> data(sagmb2011)
> sagmb2011
An object of class "qPCRset"
Size:  67 features, 55 samples
Feature types:
Feature names:           Abat Abca1 Ank ...
Feature classes:
Feature categories:      OK, Undetermined
Sample names:            CK1.Vector CK2.Vector DD2.Vector ...


> sagmb2011 <- qpcrImpute(sagmb2011, groupVars="sampleType")


> sagmb2011
An object of class "qPCRset"
Size:  67 features, 55 samples
Feature types:
Feature names:           Abat Abca1 Ank ...
Feature classes:
Feature categories:      OK, Imputed
Sample names:            CK1.Vector CK2.Vector DD2.Vector ...
```

```
> library(nondetects)
> data(sagmb2011)
> sagmb2011
An object of class "qPCRset"
Size:  67 features, 55 samples
Feature types:
Feature names:           Abat Abca1 Ank ...
Feature classes:
Feature categories:      OK, Undetermined
Sample names:            CK1.Vector CK2.Vector DD2.Vector ...


> sagmb2011 <- qpcrImpute(sagmb2011, groupVars="sampleType")


> sagmb2011
An object of class "qPCRset"
Size:  67 features, 55 samples
Feature types:
Feature names:           Abat Abca1 Ank ...
Feature classes:
Feature categories:      OK, Imputed
Sample names:            CK1.Vector CK2.Vector DD2.Vector ...
> getCtHistory(sagmb2011)
                                                 history
1                          Manually created qPCRset object.
2 qpcrImpute(object = sagmb2011, groupVars = "sampleType")
```

Are there drawbacks?

Questions about packages other than your own

Evening,

I have a txt file with 26 samples(rows) with 30 microRNA (columns) – how do I convert this into qPCRset object to do further analysis on the “nondetect” R package?

Questions about packages other than your own

Evening,

I have a txt file with 26 samples(rows) with 30 microRNA (columns) – how do I convert this into qPCRset object to do further analysis on the "nondetect" R package?

Dear Mr. McCall

I would like to use the "qpcrImpute" function on my data. My problem is that I cannot figure out how to format my data to the class qPCRset.

Relying on others to maintain software

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

ASK QUESTION | LATEST 200 | NEWS 5 | JOBS | TUTORIALS | TAGS | USERS

Showing: `htqpcr` • reset

Limit ▾ Sort ▾ Search

votes	answers	views			
0	0	13	**Error with HTqPCR rbind**	htqpcr	written 11 hours ago by NAIARA GARCÍA • 10
0	0	147	**HTqPCR: dealing with technical replicates (filterCategory) and inability to analyze datasets with filtered genes (limmaCtData).**	limma htqpcr	written 5 months ago by edu.guasch • 0
0	0	105	**A problem about changeCtLayout in HTqPCR**	htqpcr	written 5 months ago by tropicalis • 0
0	0	179	**Creation of qPCRset in R for HtqPCR**	htqpcr R bioconductor qpcrset object	written 7 months ago by mohammedtoufiq91 • 0
0	0	129	**Creating qPCRset object from Bio-Rad CFX file**	limma htqpcr	written 7 months ago by dbars • 0
0	0	158	**HTqPCR - unable to create a 'qPCRset' object**	htqpcr expressionset	written 10 months ago by jirkanov • 0
0	0	190	**Inter-run calibration in HTqPCR R package**	htqpcr R	written 12 months ago by dbarscortina • 0

Recent...

Replies

- C: dseq2 results synergy co... by steven wink • **90**
- C: DiffBind's MA plots for ... by melnuesch • **10**
- C: Mimic Counting of HTSeq-... by Martin Morgan ♦♦ **23k**
- C: Mimic Counting of HTSeq-... by Dario Strbenac • **1.4k**
- C: Cross-species RNA-seq wi... by qwzhang0601 • **0**

Votes

- How to get the final value ...
- Using the plotting function...
- A: TMM normalisation method
- A: DiffBind's MA plots for ...
- A: DiffBind's MA plots for ...

Awards • All »

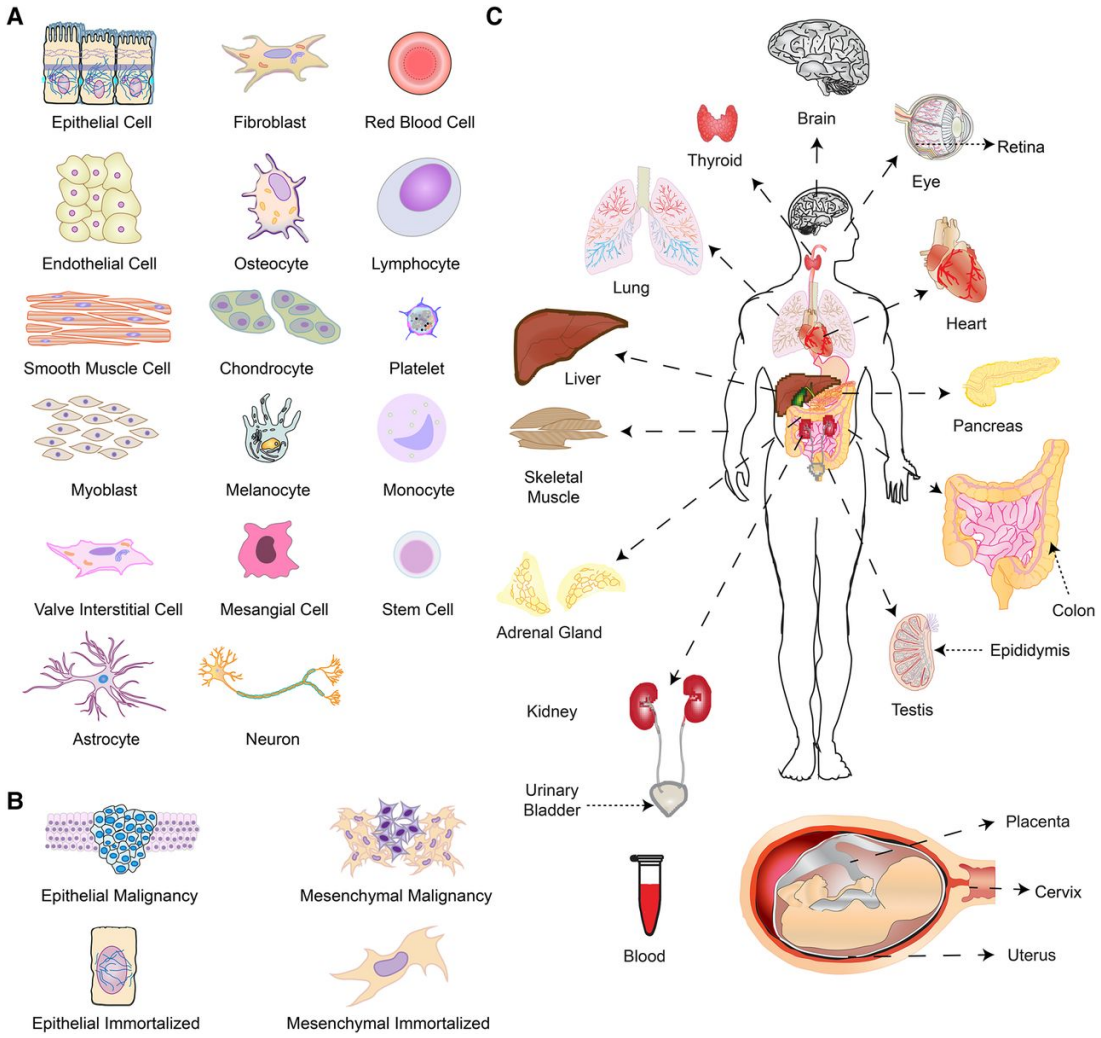
- Appreciated ♥ to shepherl ♦♦ **1.4k**
- Popular Question 👁 to shepherl ♦♦ **1.4k**
- Supporter 👍 to shepherl ♦♦ **1.4k**
- Scholar ✓ to shepherl ♦♦ **1.4k**
- Supporter 👍 to Kevin Blighe • **100**
- Teacher ☺ to Michael Love ♦ **23k**

I am not blaming the authors of HTqPCR.

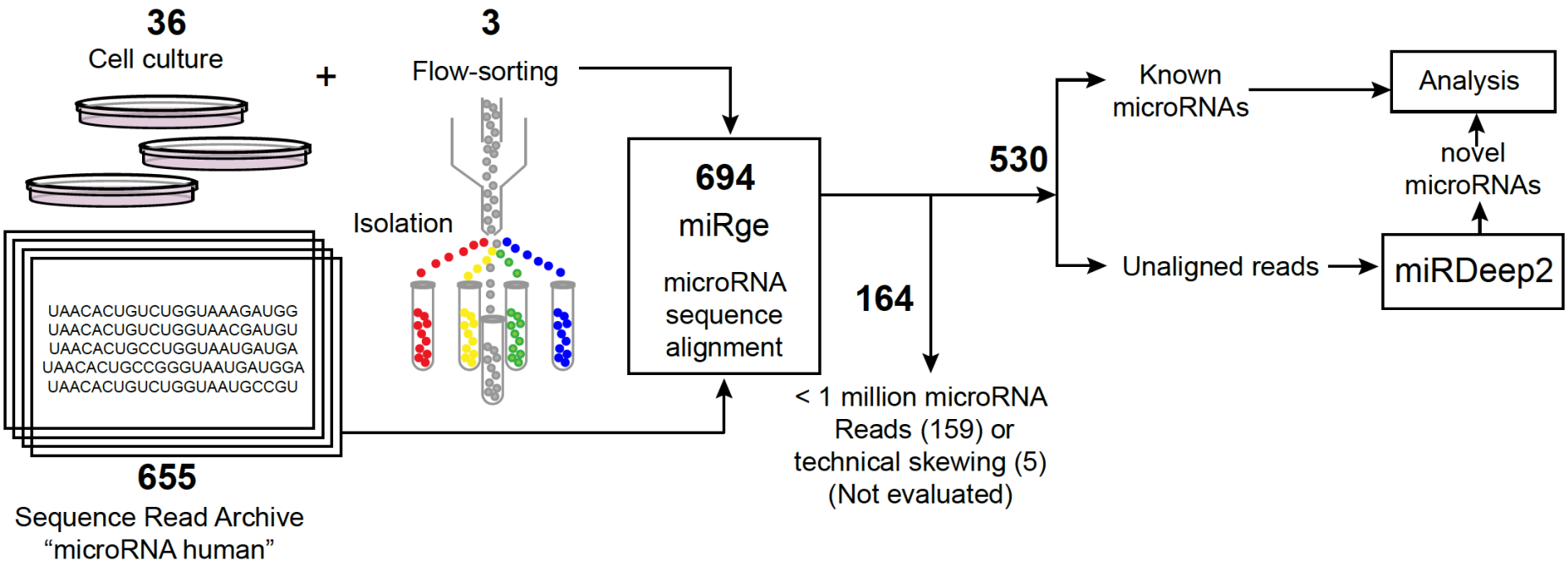
There is very little support in terms of recognition and funding for maintaining a software package.

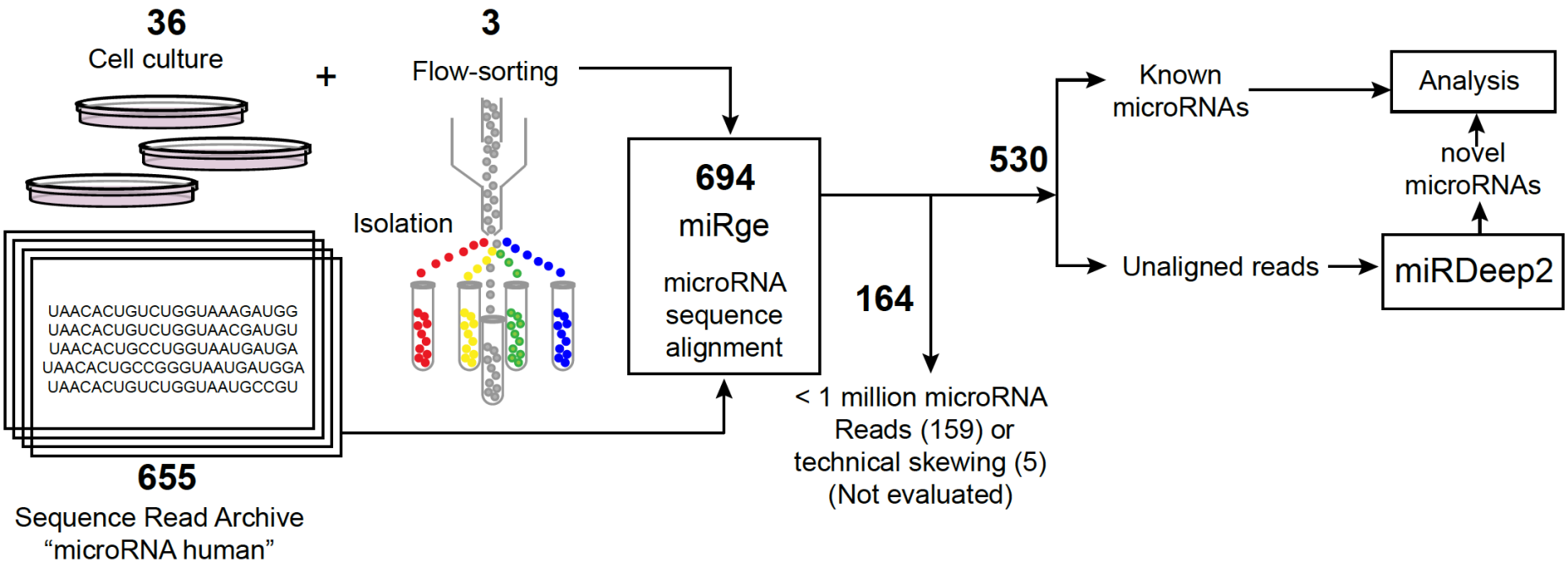
Case study #2 and a potential path forward

A generalized overview of the 530 cells and tissues included in this study.



Matthew N. McCall et al. Genome Res. 2017;27:1769-1781


36
Cell culture
+
3
Flow-sorting
Isolation
694
miRge
microRNA
sequence
alignment
530
164
< 1 million microRNA
Reads (159) or
technical skewing (5)
(Not evaluated)
Known
microRNAs
Analysis
novel
microRNAs
Unaligned reads
miRDeep2
UAACACUGUCUGGUAAAGAUGG
UAACACUGUCUGGUAACGAUGU
UAACACUGCCUGGUAAUGAUGA
UAACACUGCCGGGUAAUGAUGGA
UAACACUGUCUGGUAAUGCCGU
655
Sequence Read Archive
“microRNA human”



An integrated expression atlas of miRNAs and their promoters in human and mouse

Derek de Rie, Imad Abugessaisa […] Michiel J L de Hoon

Nature Biotechnology **35**, 872–878 (2017) | Download Citation

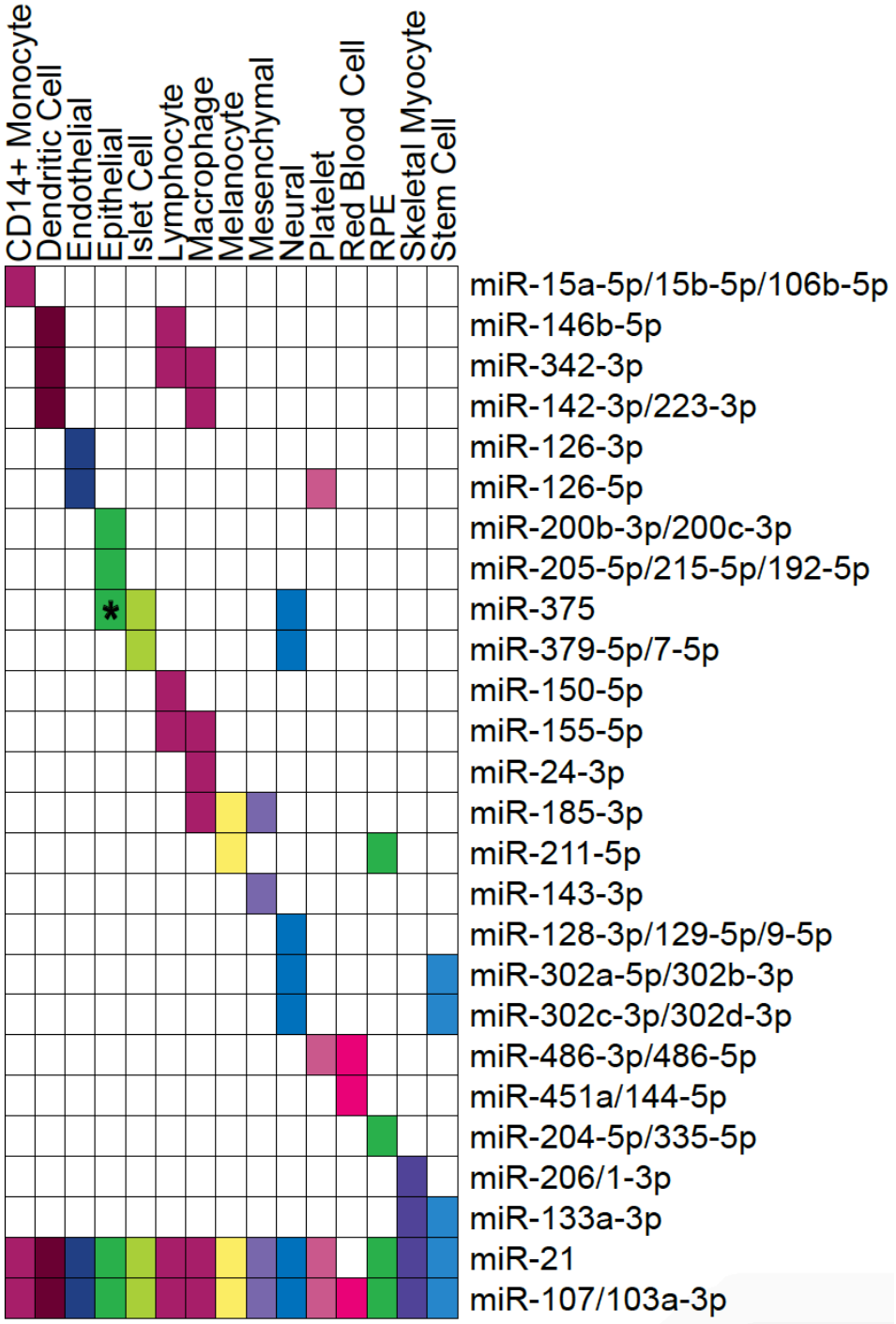
334 samples; 64 cell types

9290–9301 *Nucleic Acids Research, 2017, Vol. 45, No. 16* *Published online 11 August 2017*
doi: 10.1093/nar/gkx706

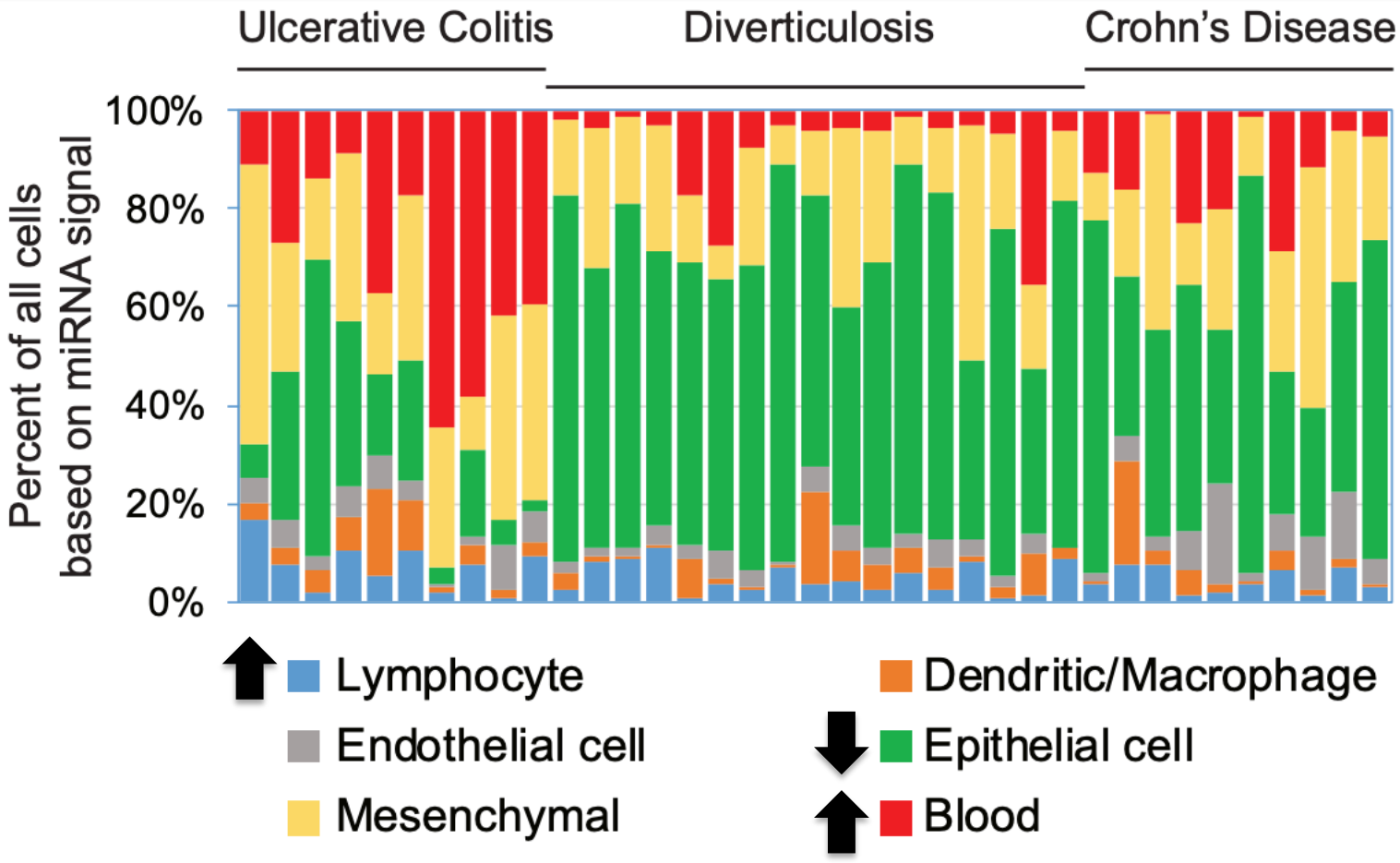
A comprehensive, cell specific microRNA catalogue of human peripheral blood

Simonas Juzenas[1,2,†], Geetha Venkatesh[1,†], Matthias Hübenthal[1,†], Marc P. Hoeppner[1], Zhipei Gracie Du[1], Maren Paulsen[1], Philip Rosenstiel[1], Philipp Senger[3], Martin Hofmann-Apitius[3], Andreas Keller[4], Limas Kupcinskas[2,5], Andre Franke[1,*] and Georg Hemmrich-Stanisak[1,*]

450 samples; 7 cell types, 3 tissues


CD14+ Monocyte
Dendritic Cell
Endothelial
Epithelial
Islet Cell
Lymphocyte
Macrophage
Melanocyte
Mesenchymal
Neural
Platelet
Red Blood Cell
RPE
Skeletal Myocyte
Stem Cell
*
miR-15a-5p/15b-5p/106b-5p
miR-146b-5p
miR-342-3p
miR-142-3p/223-3p
miR-126-3p
miR-126-5p
miR-200b-3p/200c-3p
miR-205-5p/215-5p/192-5p
miR-375
miR-379-5p/7-5p
miR-150-5p
miR-155-5p
miR-24-3p
miR-185-3p
miR-211-5p
miR-143-3p
miR-128-3p/129-5p/9-5p
miR-302a-5p/302b-3p
miR-302c-3p/302d-3p
miR-486-3p/486-5p
miR-451a/144-5p
miR-204-5p/335-5p
miR-206/1-3p
miR-133a-3p
miR-21
miR-107/103a-3p

Composition of colon tissue samples



Differences in UC vs others

How best to share these data?

SummarizedExperiment

platforms all | rank unknown | posts 5 / 0.8 / 3 / 0 | in Bioc 3.5 years

build warnings | updated before release

DOI: 10.18129/B9.bioc.SummarizedExperiment

SummarizedExperiment container

Bioconductor version: Release (3.9)

The SummarizedExperiment container contains one or more assays, each represented by a matrix-like object of numeric or other mode. The rows typically represent genomic ranges of interest and the columns represent samples.

Author: Martin Morgan, Valerie Obenchain, Jim Hester, Hervé Pagès

Maintainer: Bioconductor Package Maintainer <maintainer at bioconductor.org>

Citation (from within R, enter `citation("SummarizedExperiment")`):

> Morgan M, Obenchain V, Hester J, Pagès H (2019). *SummarizedExperiment: SummarizedExperiment container*. R package version 1.14.0.

SummarizedExperiment

platforms all | rank unknown | posts 5 / 0.8 / 3 / 0 | in Bioc 3.5 years
build warnings | updated before release

DOI: 10.18129/B9.bioc.SummarizedExperiment

SummarizedExperiment container

Bioconductor version: Release (3.9)

The SummarizedExperiment container contains one or more assays, each represented by a matrix-like object of numeric or other mode. The rows typically represent genomic ranges of interest and the columns represent samples.

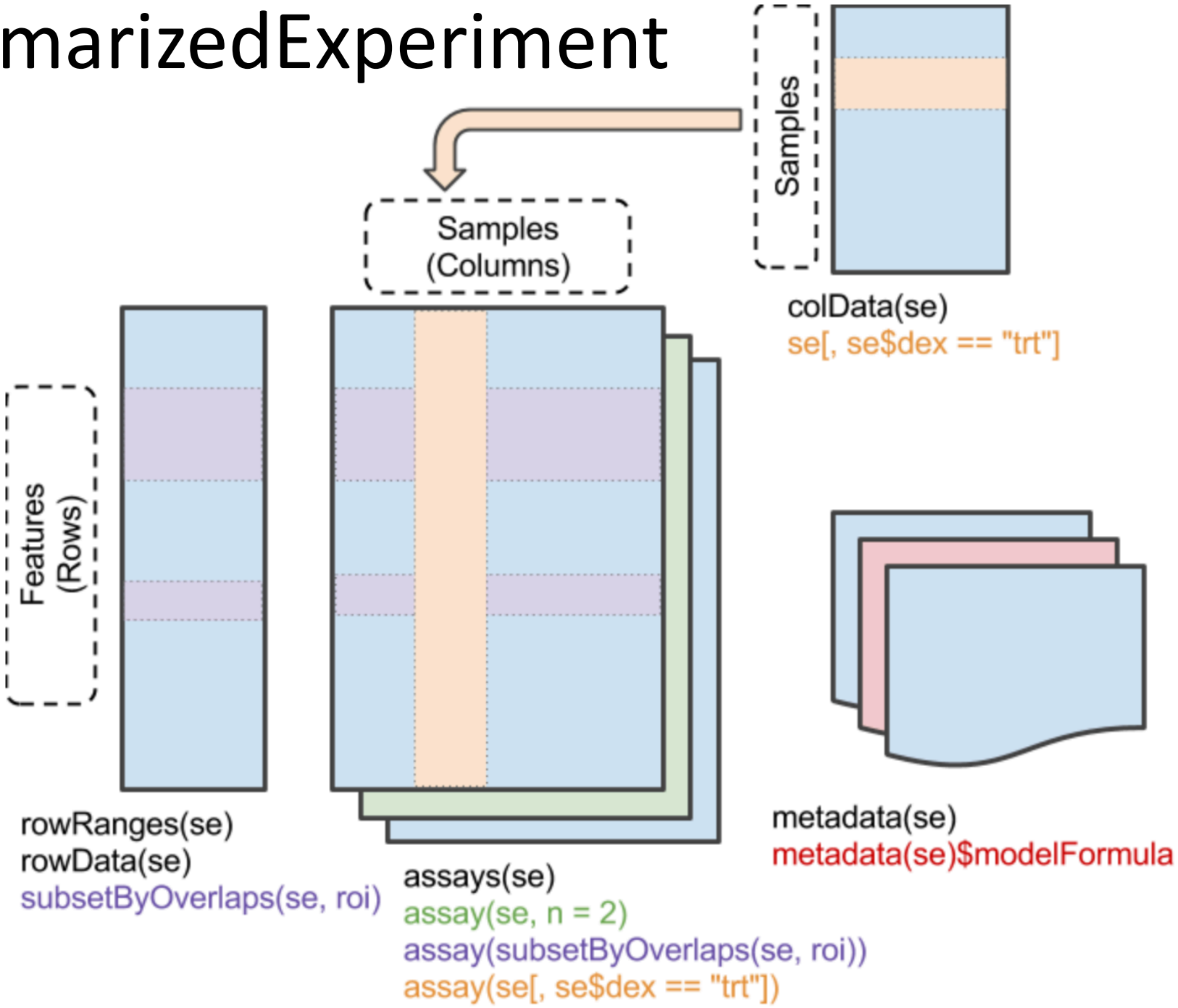
Author: Martin Morgan, Valerie Obenchain, Jim Hester, Hervé Pagès

Maintainer: Bioconductor Package Maintainer <maintainer at bioconductor.org>

Citation (from within R, enter `citation("SummarizedExperiment")`):

Morgan M, Obenchain V, Hester J, Pagès H (2019). *SummarizedExperiment: SummarizedExperiment container*. R package version 1.14.0.

SummarizedExperiment



microRNAome

platforms all | rank 301 / 371 | posts 0 | build ok

updated before release

DOI: 10.18129/B9.bioc.microRNAome

SummarizedExperiment for the microRNAome project

Bioconductor version: Release (3.9)

This package provides a SummarizedExperiment object of read counts for microRNAs across tissues, cell-types, and cancer cell-lines. The read count matrix was prepared and provided by the author of the study: Towards the human cellular microRNAome.

Author: Matthew N. McCall <mccallm at gmail.com>, Marc K. Halushka <mhalush1 at jhmi.edu>

Maintainer: Matthew N. McCall <mccallm at gmail.com>

Citation (from within R, enter `citation("microRNAome")`):

> McCall MN, Kim M, Adil M, Patil AH, Lu Y, Mitchell CJ, Leal-Rojas P, Xu J, Kumar M, Dawson VL, Dawson TM, Baras AS, Rosenberg AZ, Arking DE, Burns KH, Pandey A, Halushka M (2017). "Toward the human cellular microRNAome." *Genome Research*. doi: 10.1101/gr.222067.117, http://genome.cshlp.org/content/27/10/1769.full.pdf, http://genome.cshlp.org/content/27/10/1769.

```
> library(microRNAome)
> data("microRNAome")
> microRNAome
class: SummarizedExperiment
dim: 2546 1312
metadata(1): ''
assays(1): counts
rownames(2546): hsa-let-7a-2-3p hsa-let-7a-3p ... hsa-miR-99b-3p
  hsa-miR-99b-5p
rowData names(0):
colnames(1312): SRR2296788 ERR738403 ... SRR5756261 SRR5756262
colData names(14): sample_id organ ... sequencer flagged
```

```
> library(microRNAome)
> data("microRNAome")
> microRNAome
class: SummarizedExperiment
dim: 2546 1312
metadata(1): ''
assays(1): counts
rownames(2546): hsa-let-7a-2-3p hsa-let-7a-3p ... hsa-miR-99b-3p
  hsa-miR-99b-5p
rowData names(0):
colnames(1312): SRR2296788 ERR738403 ... SRR5756261 SRR5756262
colData names(14): sample_id organ ... sequencer flagged
> table(microRNAome$sample_category)

cancer_cell_line        cell_type           tissue
             100              791              421
> table(microRNAome$sequencer)

               AB Solid 3                AB Solid 4                        GA
                        4                         4                         8
                    GA II                    GA IIx                 HiSeq 1000
                       42                        83                        24
               HiSeq 2000 HISeq 2000 (Maybe GA IIx)                 HiSeq 2500
                      610                         9                       524
HiSeq 2500 (maybe GA IIx)               Ion Torrent                     MiSeq
                        2                         1                         1
```

Acknowledgements

- University of Rochester
  - Zachary Brehm
  - Winslow Powers
  - Qidi Yang
  - Kailey Ferger
  - Valeriia Sherina
  - Hartmut Land
  - Helene McMurray
  - Anthony Almudevar
- Johns Hopkins University
  - Marc Halushka
  - Alexander Baras
  - Avi Rosenberg
  - Dan Arking


THANK YOU