An assessment of q-PCR accuracy based on RNA-seq

presented by V. Sherina

December 14, 2015

Introduction

The **National Center for Biotechnology Information** (NCBI) is part of the United States National Library of Medicine (NLM), a branch of the National Institutes of Health.

GEO

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted.

Tools are provided to help users query and download experiments and curated gene expression profiles.

Motivation

- "A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the sequencing quality control consortium." Nature Biotechnology (2014), by SEQC/MAQC-III Consortium.
- Bioconductor package with data "RNA-seq data generated from SEQC (MAQC-III) study", by Yang Liao and Wei Shi with contributions from Steve Lianoglou.

Motivation

- "A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the sequencing quality control consortium." Nature Biotechnology (2014), by SEQC/MAQC-III Consortium.
- Bioconductor package with data "RNA-seq data generated from SEQC (MAQC-III) study", by Yang Liao and Wei Shi with contributions from Steve Lianoglou.

Paper results

The Sequencing Quality Control (SEQC) project is coordinated by the US FDA. Examining Illumina HiSeq, Life Technologies SOLiD and Roche 454 platforms at multiple laboratory sites using reference RNA samples with built-in controls, authors

- Assess RNA sequencing (RNA-seq) performance for junction discovery and differential expression profiling.
- Using complementary metrics, compare it to microarray data,

Paper results

The Sequencing Quality Control (SEQC) project is coordinated by the US FDA. Examining Illumina HiSeq, Life Technologies SOLiD and Roche 454 platforms at multiple laboratory sites using reference RNA samples with built-in controls, authors

- Assess RNA sequencing (RNA-seq) performance for junction discovery and differential expression profiling.
- Using complementary metrics, compare it to microarray data,
- and compare it to quantitative PCR (q-PCR) data.

Paper results

The Sequencing Quality Control (SEQC) project is coordinated by the US FDA. Examining Illumina HiSeq, Life Technologies SOLiD and Roche 454 platforms at multiple laboratory sites using reference RNA samples with built-in controls, authors

- Assess RNA sequencing (RNA-seq) performance for junction discovery and differential expression profiling.
- Using complementary metrics, compare it to microarray data,
- and compare it to quantitative PCR (q-PCR) data.

"Gold" standard

q-PCR-based methods:

- high sensitivity;
- and dynamic range.

"Gold" standard

q-PCR-based methods:

- high sensitivity;
- and dynamic range.
- q-PCR traditionally been used as a reference "gold" standard.

"Gold" standard

q-PCR-based methods:

- high sensitivity;
- and dynamic range.
- q-PCR traditionally been used as a reference "gold" standard.

- One challenge of q-PCR data is the presence of non-detects (those reactions failing to attain the expression threshold).

"Gold" standard

q-PCR-based methods:

- high sensitivity;
- and dynamic range.
- q-PCR traditionally been used as a reference "gold" standard.

- One challenge of q-PCR 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, recent work has shown that this introduces large biases in estimation of both absolute and differential expression.[2]

"Gold" standard

q-PCR-based methods:

- high sensitivity;
- and dynamic range.
- q-PCR traditionally been used as a reference "gold" standard.

- One challenge of q-PCR 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, recent work has shown that this introduces large biases in estimation of both absolute and differential expression.[2]
- Considerable differences in expression level measurements from different PCR-based assays can be observed.

"Gold" standard

q-PCR-based methods:

- high sensitivity;
- and dynamic range.
- q-PCR traditionally been used as a reference "gold" standard.

- One challenge of q-PCR 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, recent work has shown that this introduces large biases in estimation of both absolute and differential expression.[2]
- Considerable differences in expression level measurements from different PCR-based assays can be observed.

Ideas

- Validation has been an important part in RNA-seq publication. The differentially expressed genes (at least some) identified using RNA-seq are often validated using q-PCR.
- I am going to identify specific genes, where results of RNA-seq and q-PCR disagree.

Ideas

- Validation has been an important part in RNA-seq publication. The differentially expressed genes (at least some) identified using RNA-seq are often validated using q-PCR.
- I am going to identify specific genes, where results of RNA-seq and q-PCR disagree.
- If the disagreement caused by low detection in q-PCR experiment, it may be possible to estimate non-detected Ct values, and repeat the validation procedure.

Ideas

- Validation has been an important part in RNA-seq publication. The differentially expressed genes (at least some) identified using RNA-seq are often validated using q-PCR.
- I am going to identify specific genes, where results of RNA-seq and q-PCR disagree.
- If the disagreement caused by low detection in q-PCR experiment, it may be possible to estimate non-detected Ct values, and repeat the validation procedure.
- The claim of the project is q-PCR results may be validated using RNA-seq data.

Ideas

- Validation has been an important part in RNA-seq publication. The differentially expressed genes (at least some) identified using RNA-seq are often validated using q-PCR.
- I am going to identify specific genes, where results of RNA-seq and q-PCR disagree.
- If the disagreement caused by low detection in q-PCR experiment, it may be possible to estimate non-detected Ct values, and repeat the validation procedure.
- The claim of the project is q-PCR results may be validated using RNA-seq data.

qPCR data

- Data downloaded from GEO contained:
  - Δ Ct value 1044 by 16;

qPCR data

- Data downloaded from GEO contained:
  - Δ Ct value 1044 by 16;
  - status (Present or Absent);

qPCR data

- Data downloaded from GEO contained:
  - Δ Ct value 1044 by 16;
  - status (Present or Absent);
  - genes names.

qPCR data

- Data downloaded from GEO contained:
  - Δ Ct value 1044 by 16;
  - status (Present or Absent);
  - genes names.
- Control gene - "POLR2A";

qPCR data

- Data downloaded from GEO contained:
  - Δ Ct value 1044 by 16;
  - status (Present or Absent);
  - genes names.
- Control gene - "POLR2A";
- $Ct_{gene} = Ct_{POLR2A} - \log_2(\Delta Ct_{gene})$.

qPCR data

- Data downloaded from GEO contained:
  - Δ Ct value 1044 by 16;
  - status (Present or Absent);
  - genes names.
- Control gene - "POLR2A";
- $Ct_{gene} = Ct_{POLR2A} - \log_2(\Delta Ct_{gene})$.

Challenges: qPCR 96 well plate

Table: qPCR 96 well plates

Plate no	Plate start	Control gene	Plate end
1	1	42	96
2	97	98	192
3	193	194	288
4	289	311	384
5	385	412	480
6	481	516	576
7	577	579	672
8	673	705	768
9	769	772	864
10	865	866	960
11	961	971	1056

Therefore, data was split into 11 subsets.

Challenges: qPCR 96 well plate

Table: qPCR 96 well plates

Plate no	Plate start	Control gene	Plate end
1	1	42	96
2	97	98	192
3	193	194	288
4	289	311	384
5	385	412	480
6	481	516	576
7	577	579	672
8	673	705	768
9	769	772	864
10	865	866	960
11	961	971	1056

Therefore, data was split into 11 subsets.

Nondetects

- Nondetected signal
  - across all samples and replicates;

Nondetects

- Nondetected signal
  - across all samples and replicates;
  - nondetected in some of the replicates;

Nondetects

- Nondetected signal
  - across all samples and replicates;
  - nondetected in some of the replicates;
- I used a package "nondetects" to impute nondetected values in the second situation; and left "as is" for the first one.

Nondetects

- Nondetected signal
  - across all samples and replicates;
  - nondetected in some of the replicates;
- I used a package ”nondetects” to impute nondetected values in the second situation; and left ”as is” for the first one.
- The initial fit was estimated from all the genes.

Nondetects

- Nondetected signal
  - across all samples and replicates;
  - nondetected in some of the replicates;
- I used a package "nondetects" to impute nondetected values in the second situation; and left "as is" for the first one.
- The initial fit was estimated from all the genes.

Merge data back together

- All the data has dimension 1044 by 16 (4 trt groups, 4 replicates each);
- 118 genes were nondetected in all the replicates across samples, deleting them leads to the reduction to 926 genes.

Merge data back together

- All the data has dimension 1044 by 16 (4 trt groups, 4 replicates each);
- 118 genes were nondetected in all the replicates across samples, deleting them leads to the reduction to 926 genes.

RNA data

- RNA data from the package ”seqc” was used;
- I picked a data set, from Illumina platform;

RNA data

- RNA data from the package ”seqc” was used;
- I picked a data set, from Illumina platform;
- Counts usually refers to the number of reads that align to a particular feature. Since counts are NOT scaled by the length of the feature, all units in this category are not comparable within a sample without adjusting for the feature length.

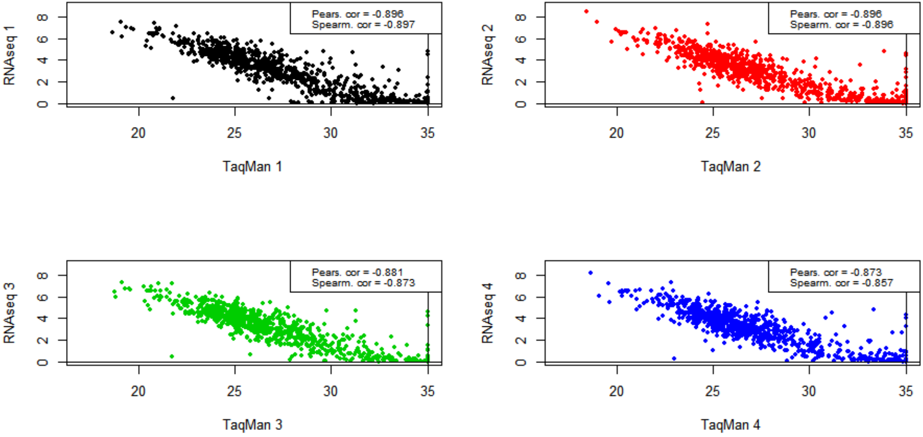
RNA data

- RNA data from the package "seqc" was used;
- I picked a data set, from Illumina platform;
- Counts usually refers to the number of reads that align to a particular feature. Since counts are NOT scaled by the length of the feature, all units in this category are not comparable within a sample without adjusting for the feature length.
- Counts per million mapped reads are counts scaled by the number of fragments that were sequenced times one million.

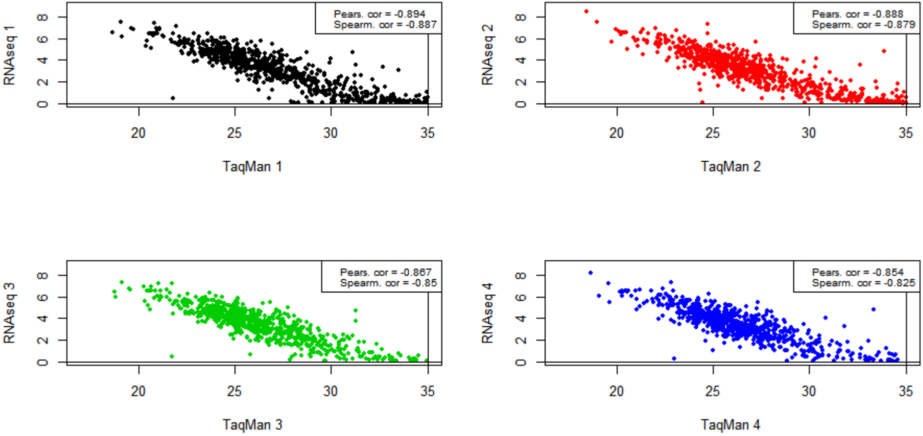
RNA data

- RNA data from the package "seqc" was used;
- I picked a data set, from Illumina platform;
- Counts usually refers to the number of reads that align to a particular feature. Since counts are NOT scaled by the length of the feature, all units in this category are not comparable within a sample without adjusting for the feature length.
- Counts per million mapped reads are counts scaled by the number of fragments that were sequenced times one million.

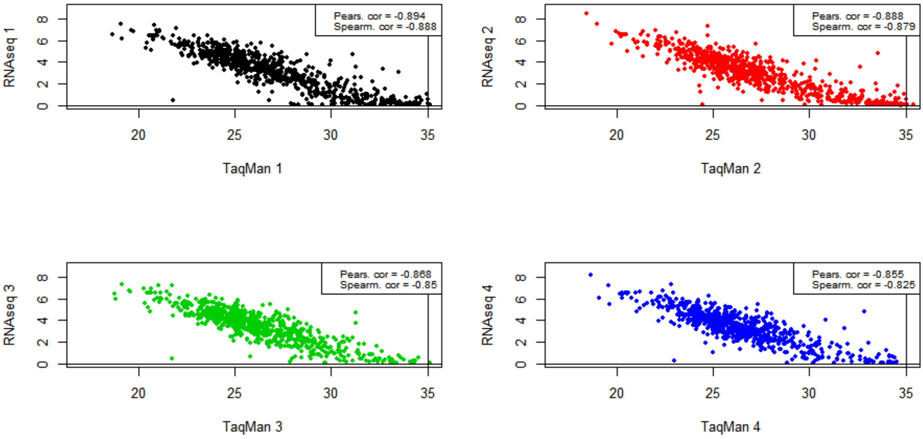
All genes in the data



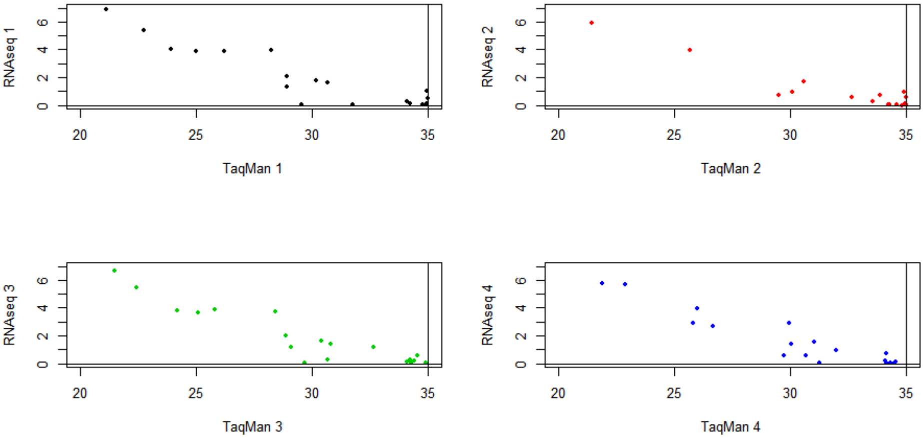
Genes absent across samples were excluded



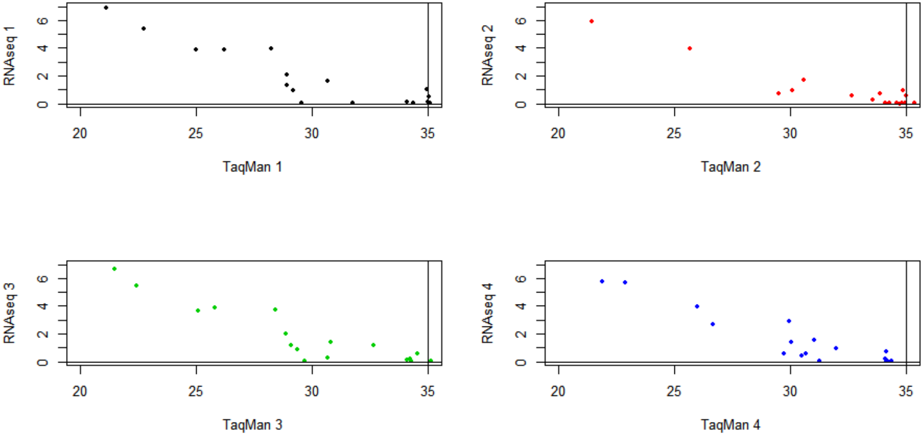
With imputed missing values



Extreme values, excluded genes



Extreme values, imputed missing data points



Absent in RNAseq data genes

- CRISP3 (Cysteine-Rich Secretory Protein 3) is a Protein Coding gene. Diseases associated with CRISP3 include prostate cancer.
- A gene, CYP2B6, encodes a member of the cytochrome P450 superfamily of enzymes. The cytochrome P450 proteins are monooxygenases which catalyze many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids.

Absent in RNAseq data genes

- CRISP3 (Cysteine-Rich Secretory Protein 3) is a Protein Coding gene. Diseases associated with CRISP3 include prostate cancer.
- A gene, CYP2B6, encodes a member of the cytochrome P450 superfamily of enzymes. The cytochrome P450 proteins are monooxygenases which catalyze many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids.
- RNF186 (Ring Finger Protein 186) is a Protein Coding gene. Diseases associated with RNF186 include ulcerative colitis.

Absent in RNAseq data genes

- CRISP3 (Cysteine-Rich Secretory Protein 3) is a Protein Coding gene. Diseases associated with CRISP3 include prostate cancer.
- A gene, CYP2B6, encodes a member of the cytochrome P450 superfamily of enzymes. The cytochrome P450 proteins are monooxygenases which catalyze many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids.
- RNF186 (Ring Finger Protein 186) is a Protein Coding gene. Diseases associated with RNF186 include ulcerative colitis.
- MIA2 (Melanoma Inhibitory Activity 2) is a Protein Coding gene. May play a role in the pathophysiology of liver disease and may serve as a marker of liver damage.

Absent in RNAseq data genes

- CRISP3 (Cysteine-Rich Secretory Protein 3) is a Protein Coding gene. Diseases associated with CRISP3 include prostate cancer.
- A gene, CYP2B6, encodes a member of the cytochrome P450 superfamily of enzymes. The cytochrome P450 proteins are monooxygenases which catalyze many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids.
- RNF186 (Ring Finger Protein 186) is a Protein Coding gene. Diseases associated with RNF186 include ulcerative colitis.
- MIA2 (Melanoma Inhibitory Activity 2) is a Protein Coding gene. May play a role in the pathophysiology of liver disease and may serve as a marker of liver damage.

Absent in qPCR data genes

- ABCC11 (sample A) and ABCB11 (sample B) gene. The proteins encoded by these genes are members of the superfamily of ATP-binding cassette (ABC) transporters. ABC proteins transport various molecules across extra- and intra-cellular membranes.
- AFP Gene (sample D). This gene encodes alpha-fetoprotein, a major plasma protein produced by the yolk sac and the liver during fetal life. Alpha-fetoprotein expression in adults is often associated with hepatoma or teratoma. However, hereditary persistance of alpha-fetoprotein may also be found in individuals with no obvious pathology.

Absent in qPCR data genes

- ABCC11 (sample A) and ABCB11 (sample B) gene. The proteins encoded by these genes are members of the superfamily of ATP-binding cassette (ABC) transporters. ABC proteins transport various molecules across extra- and intra-cellular membranes.
- AFP Gene (sample D). This gene encodes alpha-fetoprotein, a major plasma protein produced by the yolk sac and the liver during fetal life. Alpha-fetoprotein expression in adults is often associated with hepatoma or teratoma. However, hereditary persistance of alpha-fetoprotein may also be found in individuals with no obvious pathology.

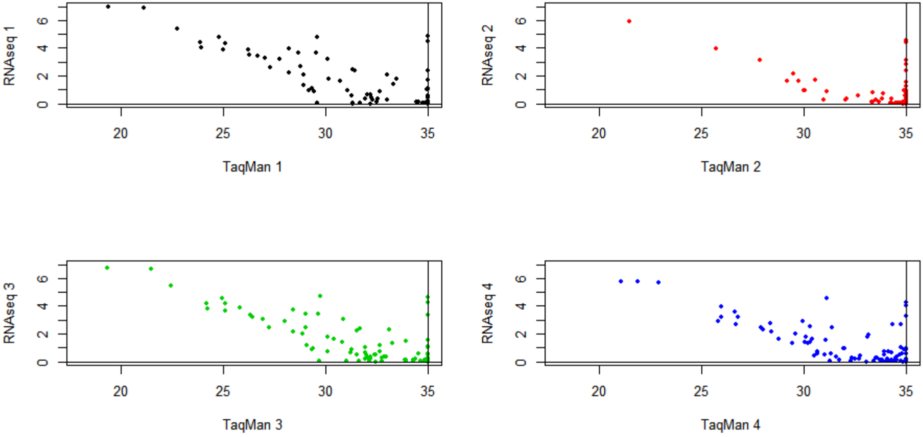
The information about genes was taken from "GeneCards".

Absent in qPCR data genes

- ABCC11 (sample A) and ABCB11 (sample B) gene. The proteins encoded by these genes are members of the superfamily of ATP-binding cassette (ABC) transporters. ABC proteins transport various molecules across extra- and intra-cellular membranes.
- AFP Gene (sample D). This gene encodes alpha-fetoprotein, a major plasma protein produced by the yolk sac and the liver during fetal life. Alpha-fetoprotein expression in adults is often associated with hepatoma or teratoma. However, hereditary persistance of alpha-fetoprotein may also be found in individuals with no obvious pathology.

The information about genes was taken from "GeneCards".

Extreme values, all genes



Results

- Several not matching genes were identified during this project.
- Correlation between RNAseq and qPCR results found to be similar to the results from the paper [1].

Results

- Several not matching genes were identified during this project.
- Correlation between RNAseq and qPCR results found to be similar to the results from the paper [1].
- The R package "TaqManRNAseq" was created.

Results

- Several not matching genes were identified during this project.
- Correlation between RNAseq and qPCR results found to be similar to the results from the paper [1].
- The R package "TaqManRNAseq" was created.
- It contains total of 5 functions.

Results

- Several not matching genes were identified during this project.
- Correlation between RNAseq and qPCR results found to be similar to the results from the paper [1].
- The R package "TaqManRNAseq" was created.
- It contains total of 5 functions.
- This project gives a motivation for future research in this area.

Results

- Several not matching genes were identified during this project.
- Correlation between RNAseq and qPCR results found to be similar to the results from the paper [1].
- The R package "TaqManRNAseq" was created.
- It contains total of 5 functions.
- This project gives a motivation for future research in this area.

References

- SEQC/MAQC-III Consortium *A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium* Nature Biotechnology 32, 903914, 2014.
- M.N. McCall, H.R. McMurray, H. Land and A. Almudevar *On non-detects in Quantitative real-time PCR data.* Bioinformatics V. 30 no. 16, 2310-2316, 2014.
- GeneCards: *The Human Gene Database*, http://www.genecards.org/

Thank you for your attention!

Questions?

Thank you for your attention!
Questions?