Birytkov (40)

Statistics for Genomic Data Analysis, Final Data Analysis Reports

- give informative title

2 **Background/Introduction 5%** - samples not probes
- which chip / # probe sets

6 **Quality assessment 15%**
- Don't need numerical NUSE/RLE summaries
+ Define all terms

/8 Description of quality assessment 0/8
/3 Excluded poor quality slides (if relevant) 3/3
(*) /4 Graphical support for decisions - square pseudo images
- Don't need Figures 3/4/5/6
3/4
- horizontal line at 1.05 NUSE

2 **Normalization 10%**

2/6 Description of normalization/expression measure used . mathematically
0/2 Graphical support for decisions
0/2 Appropriateness of normalization

10 **Statistical analyses: Differential Expression 25%** Don't need Figure 9a

0/5 Description of model (including design matrix, contrasts if necessary)
0/5 Multiple testing/adjusted p-values given - explain (t mod t)
2/5 Appropriate basis for gene ranking (and explanation)
5/5 Estimated number of differentially expressed genes (and explanation)
3/5 Volcano plot, other appropriate plots - explain

5 (**) **Statistical analyses: Cluster Analysis 20%**

0/10 Description of clustering algorithm used not clear / not done
3/5 Relevant graphics - just choose 1 method and describe completely
2/5 Interpretation of resulting clusters not clear

5 **Conclusions 10%** - recap analysis
- somewhat vague

Other

4 **Gene list 5%** - make 'pretty' columns
- separate page at end, not in text

5 **Reproducible R code 5%** - in separate text file

1 **Overall Presentation 5%**
- use 12 point size
- Display all chips in one figure
- no raw R output/input
- plots not explained
- small plot label size
- number all tables

(*) images don't confirm hypothesis → support

(**) Ward's is for joining clusters, 1-corr is for joining objects

Corbelli 49

Statistics for Genomic Data Analysis, Final Data Analysis Reports

5 **Background/Introduction 5%** (probe sets / 'genes') RMA values (not extracted values)

2 **Quality assessment 15%**

0/8 Description of quality assessment
1/3 Excluded poor quality slides (if relevant)
1/4 Graphical support for decisions

not done/not clear

3 **Normalization 10%**

0/6 Description of normalization/expression measure used
2/2 Graphical support for decisions
1/2 Appropriateness of normalization

16 **Statistical analyses: Differential Expression 25%**

5/5 Description of model (including design matrix, contrasts if necessary)
(*) 2/5 Multiple testing/adjusted p-values given – Choose 1 and explain (mod t)
1/5 Appropriate basis for gene ranking (and explanation)
5/5 Estimated number of differentially expressed genes (and explanation) (clarify)
3/5 Volcano plot, other appropriate plots – $\log_{10}$ adj p not B

9 **Statistical analyses: Cluster Analysis 20%**

2/10 Description of clustering algorithm used vague + imprecise – I should be able to replicate what you did
4/5 Relevant graphics
3/5 Interpretation of resulting clusters

4 **Conclusions 10%** – recap analysis
– Vague

Other

3 **Gene list 5%** – make 'pretty' column labels

5 (**Reproducible R code 5%**) – Did you take coef = 2 in toptable?

2 **Overall Presentation 5%**
– a lot of blank space
– Don't need Figure 3

(*) B stat does not have p-values – use mod t adj p to rank

El Messaoudi 68

Statistics for Genomic Data Analysis, Final Data Analysis Reports

- give informative title

3 **Background/Introduction 5%** samples (not 'CEL files')
- which chip / # probe sets

8 **Quality assessment 15%** not 'ensure'

2/8 Description of quality assessment Define all terms
3/3 Excluded poor quality slides (if relevant) 1.05 not 1.1
3/4 Graphical support for decisions (weights not resids)

Normalization 10% Figure 3 unnecessary

6
2/6 Description of normalization/expression measure used mathematically
2/2 Graphical support for decisions
2/2 Appropriateness of normalization

Statistical analyses: Differential Expression 25%

17
5/5 Description of model (including design matrix, contrasts if necessary)
3/5 Multiple testing/adjusted p-values given
⊛ 2/5 Appropriate basis for gene ranking (and explanation) (|mod t|)
4/5 Estimated number of differentially expressed genes (and explanation)
3/5 Volcano plot, other appropriate plots - -log10 adj p

Statistical analyses: Cluster Analysis 20%

11
7/10 Description of clustering algorithm used - use Ward instead of avg linkage
2/5 Relevant graphics heatmap
2/5 Interpretation of resulting clusters (not clear)

10 **Conclusions 10%** good

Other

4 **Gene list 5%** - make 'pretty' column headings

5 **Reproducible R code 5%** - separate text file

4 **Overall Presentation 5%**
- name chips PD, etc, not PO.CEL
- log2 intensities (not 'raw') + are they PM or both?
- small plot label size
- Don't need Figure 6

⊛ adj P corresponds to mod t (not B)
⇒ Don't need Bstat, just rank on adj P
(|mod t|)

Hengl

75

Statistics for Genomic Data Analysis, Final Data Analysis Reports

- give informative title
\# probe sets

4 **Background/Introduction 5%**

Quality assessment 15%

15

Description of quality assessment
Excluded poor quality slides (if relevant)
Graphical support for decisions

Normalization 10%

8
6/6 Description of normalization/expression measure used
0/2 Graphical support for decisions
2/2 Appropriateness of normalization

12 **Statistical analyses: Differential Expression 25%**

5/5 Description of model (including design matrix, contrasts if necessary)
1/5 Multiple testing/adjusted p-values given - explain (+ mod t)
2/5 Appropriate basis for gene ranking (and explanation) - not clear -
0/5 Estimated number of differentially expressed genes (and explanation)
4/5 Volcano plot, other appropriate plots - explain

11 **Statistical analyses: Cluster Analysis 20%**

6/10 Description of clustering algorithm used - incomplete
5/5 Relevant graphics - use 1-corr not Manhattan
0/5 Interpretation of resulting clusters

10 **Conclusions 10%** (ok)

Other

5 **Gene list 5%**

5 **Reproducible R code 5%**

5 **Overall Presentation 5%**

- label chips P0, ... not P0.CEL
- no refs

Ho

53

Statistics for Genomic Data Analysis, Final Data Analysis Reports

3 - give informative title

Background/Introduction 5% - # probesets - which chip

Quality assessment 15%

7
- 2/8 Description of quality assessment Figure 2? Define all terms
- 3/3 Excluded poor quality slides (if relevant) incomplete
- 2/4 Graphical support for decisions - pseudo-images (not pseudo-chip)

add horizontal line at 1.05 NUSE

Normalization 10%

6
- 2/6 Description of normalization/expression measure used mathematically
- 2/2 Graphical support for decisions
- 2/2 Appropriateness of normalization

Statistical analyses: Differential Expression 25% - describe methods not R details

7
- 0/5 Description of model (including design matrix, contrasts if necessary) mathematically
- 1/5 Multiple testing/adjusted p-values given - explanation (+ mod t)
- 1/5 Appropriate basis for gene ranking (and explanation)
- 3/5 Estimated number of differentially expressed genes (and explanation)
- 2/5 Volcano plot, other appropriate plots Why B-stat? You used adj p

Statistical analyses: Cluster Analysis 20%

12
- 4/10 Description of clustering algorithm used incomplete Use Ward + 1-corr
- 5/5 Relevant graphics
- 3/5 Interpretation of resulting clusters not 'significantly' - no statistical test - interpret colors

7 **Conclusions 10%** (recap) - use paragraphing

Other

3 **Gene list 5%** - too many digits - make 'pretty' column headings

5 **Reproducible R code 5%**

3 **Overall Presentation 5%**

(use 12 point size)
- label chips P0, ... not P0.CEL
- $\log_2$ PM (not 'raw')
- small plot label size
- cite primary refs
- incomplete refs
- Figure not Fig.

46

Hyun

Statistics for Genomic Data Analysis, Final Data Analysis Reports

-give informative title

3 **Background/Introduction 5%** - which chip / # probe sets

Quality assessment 15%

7

2/8 Description of quality assessment — methods, not package/function details
Define all terms

3/3 Excluded poor quality slides (if relevant)

2/4 Graphical support for decisions — all chips on same plot
- horizontal line at 1.05 NUSE
square pseudo-images weights

6 **Normalization 10%**

2/6 Description of normalization/expression measure used — mathematical(only)

2/2 Graphical support for decisions

2/2 Appropriateness of normalization

Statistical analyses: Differential Expression 25%

- don't need R details

10

4/5 Description of model (including design matrix, contrasts if necessary) (+mod t)

2/5 Multiple testing/adjusted p-values given — vague + define all terms; not 'ensure low chance'

3/5 Appropriate basis for gene ranking (and explanation) ⊛

0/5 Estimated number of differentially expressed genes (and explanation)

2/5 Volcano plot, other appropriate plots

12 **Statistical analyses: Cluster Analysis 20%**

incomplete

5/10 Description of clustering algorithm used Combine cluster method = ? (use Ward's)

5/5 Relevant graphics

2/5 Interpretation of resulting clusters over-interpreted

'robust' = ??

8 **Conclusions 10%**
(vague in parts)

Other

- make 'pretty' column headings

2 **Gene list 5%** - too many digits
- not 0.000 - use scientific notation

5 **Reproducible R code 5%**

3 **Overall Presentation 5%**

- label chips P0, ..., not P0.CEL
- small label size on plots
- also cite Irizarry et al. NAR
- Don't need Figure 6

⊛ Why show B-stat when you rank based on adj p?

Lavrenko 38

Statistics for Genomic Data Analysis, Final Data Analysis Reports

- give informative titles

3 **Background/Introduction 5%** - don't need .CEL file detail

- which chip/# probe sets

Quality assessment 15%

(*)

2/8 Description of quality assessment mathematically + define all terms

3/3 Excluded poor quality slides (if relevant) horizontal line at 1.05 NUSE

2/4 Graphical support for decisions - weight pseudo-images

Figures 1+2 are useless

0 **Normalization 10%** not done?

mathematically

0/6 Description of normalization/expression measure used

0/2 Graphical support for decisions

0/2 Appropriateness of normalization

8 **Statistical analyses: Differential Expression 25%**

mathematically

1/5 Description of model (including design matrix, contrasts if necessary)

1/5 Multiple testing/adjusted p-values given (t mod t) → use BH for FDR + EXPLAIN

0/5 Appropriate basis for gene ranking (and explanation) ?

5/5 Estimated number of differentially expressed genes (and explanation)

1/5 Volcano plot, other appropriate plots - why p-stat

1+ explain

Testis ONLY

10 **Statistical analyses: Cluster Analysis 20%**

incomplete - choose 1 method and describe completely (use Ward's + 1-corr)

4/10 Description of clustering algorithm used

5/5 Relevant graphics

1/5 Interpretation of resulting clusters

7 **Conclusions 10%** - over-stated - validates?

+ Don't need table 3

Other

3 (**) **Gene list 5%** - incomplete + make pretty labels

5 (**Reproducible R code 5%**) - in a separate text file

2 **Overall Presentation 5%**

(*) all chips in SAME graph, not separated by type

- small plot labels
- name chips P0, ..., not P0.CEL
- no refs
- plots support, not 'confirm'

(**) Table incorrect, you did not exclude P0, T0

- no raw R input/output

Monaco 64

Statistics for Genomic Data Analysis, Final Data Analysis Reports

3 **Background/Introduction 5%** – which chips? – # probe sets? a little incomplete

9 **Quality assessment 15%** – model over parametrized as written; bounds on i, j, k – Define all quantities

5/8 Description of quality assessment
1/3 Excluded poor quality slides (if relevant)
3/4 Graphical support for decisions – horizontal line 1.05 NUSE – color | square pseudo images

8 **Normalization 10%**

not clear 5/6 Description of normalization/expression measure used
1/2 Graphical support for decisions – Boxplots more informative here
2/2 Appropriateness of normalization

17 **Statistical analyses: Differential Expression 25%**

3/5 Description of model (including design matrix, contrasts if necessary) – methods not R commands
2/5 Multiple testing/adjusted p-values given – explain + define all terms (+ mod t)
3/5 ⊛ Appropriate basis for gene ranking (and explanation)
5/5 Estimated number of differentially expressed genes (and explanation)
4/5 Volcano plot, other appropriate plots

14 **Statistical analyses: Cluster Analysis 20%**

7/10 Description of clustering algorithm used – cluster samples (not genes)
5/5 Relevant graphics
2/5 Interpretation of resulting clusters – not entirely clear

Conclusions 10% – recap analyses

Other

3 **Gene list 5%** – top 50 in single table – make 'pretty' table columns

5 (**Reproducible R code 5%**)

5 **Overall Presentation 5%**

– Don't need resumé, contents
– not 'confirm poor quality' (indicate)
– refs incomplete, cite primary refs

⊛ why B-stat? adj p goes with mod t
– Don't need table of figures

Pelletier 55

Statistics for Genomic Data Analysis, Final Data Analysis Reports

3 **Background/Introduction 5%** – which chip/# probe sets

7 **Quality assessment 15%**

2/2 Description of quality assessment mathematically + define all terms
0/3 Excluded poor quality slides (if relevant) – unclear + horizontal line
2/4 Graphical support for decisions – weight pseudo-images at 1.05 NUSE

6 **Normalization 10%**

mathematical
2/6 Description of normalization/expression measure used
2/2 Graphical support for decisions
2/2 Appropriateness of normalization

10 **Statistical analyses: Differential Expression 25%**

mathematical
1/5 Description of model (including design matrix, contrasts if necessary)
1/5 Multiple testing/adjusted p-values given – complete explanation (define all (+ mod t) terms)
4/5 Appropriate basis for gene ranking (and explanation)
0/5 Estimated number of differentially expressed genes (and explanation)
4/5 Volcano plot, other appropriate plots – $-\log_{10}$ adj p

13 **Statistical analyses: Cluster Analysis 20%**

– incomplete
4/10 Description of clustering algorithm used – use Ward's with 1-cor
5/5 Relevant graphics
4/5 Interpretation of resulting clusters

9 **Conclusions 10%** – # DE genes

Other – Don't need Table 1

3 **Gene list 5%** – all on 1 page / too many digits / 'pretty' column names

5 **Reproducible R code 5%**

2 **Overall Presentation 5%**

- Don't need Contents
- small plot labels
- incomplete and superficial
- Number all Figures
- Don't need MA plot

Pizzichini

48

Statistics for Genomic Data Analysis, Final Data Analysis Reports

4 **Background/Introduction 5%** (samples not CEL files) – which chips (probe sets/'genes')

Quality assessment 15%

7

2/8 Description of quality assessment – methods not R details + mathematical description

2/3 Excluded poor quality slides (if relevant)

3/4 Graphical support for decisions – plot all chips in same plots – horizontal line at 1.05 on NUSE plot – not 'confirms'

Normalization 10%

5

3/6 Description of normalization/expression measure used – use paragraphing, description is hard to read

1/2 Graphical support for decisions – pre + post-norm boxplots

1/2 Appropriateness of normalization

Statistical analyses: Differential Expression 25%

9

4/5 Description of model (including design matrix, contrasts if necessary) – Don't need R details

2/5 Multiple testing/adjusted p-values given – confusing and vague explanation (t mod t)

1/5 Appropriate basis for gene ranking (and explanation) – why B-stat – use mod t (adj) p-value

0/5 Estimated number of differentially expressed genes (and explanation)

2/5 Volcano plot, other appropriate plots – incorrect

Statistical analyses: Cluster Analysis 20%

12

4/10 Description of clustering algorithm used – somewhat incomplete – use Ward not FN – 1-corr not Euclidean

4/5 Relevant graphics

4/5 Interpretation of resulting clusters – cluster samples (only) not genes

3 **Conclusions 10%**

– recap before summarizing findings

– confusing statements

Other

2 **Gene list 5%** – incorrect – too many digits – make 'pretty' column headings

5 (**Reproducible R code 5%**)

1 **Overall Presentation 5%**

– label chips P0, ... not P0.CEL

– small labels on plots

→ no Raw R (input or output)

– Don't need MA plots

– no refs

(67)

Rais

Statistics for Genomic Data Analysis, Final Data Analysis Reports

-give informative title -which chip

-number of probe sets

3 **Background/Introduction 5%**

Quality assessment 15%

mathematically + ~~De~~fine all terms (+explain)

7 2/8 Description of quality assessment
2/3 Excluded poor quality slides (if relevant)
3/4 Graphical support for decisions - add horizontal line at 1.05 NUSE (not 1.10
-square weights

5 **Normalization 10%**

mathematically

2/6 Description of normalization/expression measure used
1/2 Graphical support for decisions - pre/post boxplots; don't need post density
2/2 Appropriateness of normalization

22 **Statistical analyses: Differential Expression 25%** good

5/5 Description of model (including design matrix, contrasts if necessary)
3/5 Multiple testing/adjusted p-values given + mod t
5/5 Appropriate basis for gene ranking (and explanation)
5/5 Estimated number of differentially expressed genes (and explanation)
4/5 Volcano plot, other appropriate plots - Put explanation/description before plot then interpretation after

Statistical analyses: Cluster Analysis 20%

12 complete

5/10 Description of clustering algorithm used what distance? which genes?
5/5 Relevant graphics
2/5 Interpretation of resulting clusters

6 **Conclusions 10%** not 'confirms'
-recap analyses

Other

4 **Gene list 5%** make 'pretty' column headings

5 **Reproducible R code 5%**

3 **Overall Presentation 5%**

-name chips P0, ... not P0.CEL
-number all plots
-Don't need MA plot
-also cite Smyth
-no raw R input/output

Stoll 41

Statistics for Genomic Data Analysis, Final Data Analysis Reports

3 **Background/Introduction 5%** which chip / #probe sets

4 **Quality assessment 15%**

2/4 Description of quality assessment — Describe methods + define all terms
0/3 Excluded poor quality slides (if relevant) — incorrect interpretation of NUSE plot
2/4 Graphical support for decisions — (weights pseudo-images not done

0 **Normalization 10%**

Description of normalization/expression measure used
Graphical support for decisions
Appropriateness of normalization

10 **Statistical analyses: Differential Expression 25%** Describe methods not R details

1/5 Description of model (including design matrix, contrasts if necessary) — mathematically
2/5 Multiple testing/adjusted p-values given — define + explain (+ mod t)
4/5 Appropriate basis for gene ranking (and explanation)
0/5 Estimated number of differentially expressed genes (and explanation)
3/5 Volcano plot, other appropriate plots — illegible

9 **Statistical analyses: Cluster Analysis 20%**

3/10 Description of clustering algorithm used — explain — use Ward not complete
5/5 Relevant graphics — 1-corr not Euclidean
1/5 Interpretation of resulting clusters

5 **Conclusions 10%**

- First recap
- vague + some incorrect statements

Other

3 **Gene list 5%**

- Dont need Table 1
- too many digits
- make 'pretty' column labels

(table incorrect because didnt exclude P0, T0

5 **Reproducible R code 5%**

2 **Overall Presentation 5%**

- use 12 point size
- name chips P0, ..., not P0.CEL...
- number all plots
- too much blank space
- no refs

Zhang 54

Statistics for Genomic Data Analysis, Final Data Analysis Reports

3 - give informative title

Background/Introduction 5% - which chip / # probe sets

- samples not CEL files

9 **Quality assessment 15%**

4/8 Description of quality assessment — complete mathematical description + Define all terms

3/3 Excluded poor quality slides (if relevant) — horizontal line at 1.05 NUSE

2/4 Graphical support for decisions — weights (all chips) | square shape (not resids)

4 **Normalization 10%** log2 — complete mathematical

3/6 Description of normalization/expression measure used

0/2 Graphical support for decisions

1/2 Appropriateness of normalization

15 **Statistical analyses: Differential Expression 25%**

1/5 Description of model (including design matrix, contrasts if necessary) = ?

3/5 * Multiple testing/adjusted p-values given (t mod t)

3/5 Appropriate basis for gene ranking (and explanation)

5/5 Estimated number of differentially expressed genes (and explanation) — here as well as conclusion

3/5 Volcano plot, other appropriate plots — use adj P, not B (or mod t)

13 **Statistical analyses: Cluster Analysis 20%**

5/10 Description of clustering algorithm used (complete) (ok, but in exam you will need to cluster samples) Don't need R details, use 1-corr not Euclidean

3/5 Relevant graphics silhouette

5/5 Interpretation of resulting clusters (ok) (silhouette not defined)

Conclusions 10%

Other

3 **Gene list 5%** - make 'pretty' column headings

5 **Reproducible R code 5%**

2 **Overall Presentation 5%**

- use 12 point size
- also cite Irizarry et al. (NAR)
- small label size
- Did you do toptable on 2^ND parameter?

* p-values correspond to mod t not B-stat