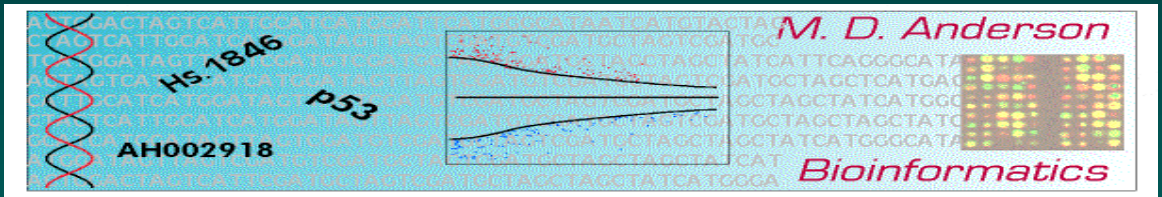
Writing Good Reports

Keith A. Baggerly
Bioinformatics and Computational Biology
UT M. D. Anderson Cancer Center
kabagg@mdanderson.org

SISBID, July 20, 2016



Writing Good Reports Enhances Reproducibility

Why are we treating report writing as a separate topic?

In part, based on negative feedback we got.

Early on in our interchanges with the Duke group, we started experimenting with literate programming and *Sweave*.

We did this because we were frustrated by our inability to get straight answers to what we thought were reasonable questions, and thought it might help to be able to say

“Here’s exactly what we’re doing. Where are we wrong?”

Code is Only Part of the Story

As we found more mistakes, we grew more convinced of the importance of documentation.

We could easily see ourselves making similar mistakes.

Based on this conviction, we started requiring our analysts to generate similarly detailed reports.

Our collaborators hated it.

Huh?

After a few discussions, it became clear that there were no objections to tracking computations or reproducibility *per se*.

Rather, they had objections because there was now so much emphasis on code.

They couldn’t quickly find the results they cared about from the computer output.

So, we refocused on how we could restructure the reports to meet both goals.

What would our audience find most comprehensible?

What Should A Report Contain?

At the outset, a report should clearly state

The underlying biological question we hope to address.

What experiments were performed, and how these experiments are expected to provide some type of answer to the question.

What analyses are being performed in the report at hand.

The results of these analyses.

The conclusions we draw from these results, and indicate where we should go next.

It should then present the analysis itself.

The Executive Summary: Structure

Introduction
– Background and Rationale
– Objectives
Data & Methods
– Data Description
– Statistical/Analytical Methods
Results
Conclusions

This what our clinicians are used to, and find easy to digest.

This summary should be in prose, ideally ≤ 2 pages.

The Importance of Communication

Some stuff I do may seem alien to my collaborators.

The converse occasionally applies as well.

I know I can analyze data better when I understand the data's limitations, and I can likewise think of better approaches if I have a clear understanding of the end goal.

My collaborators know a lot about the data that I don't, and they can make better suggestions to me if they understand what I'm planning to try, and why.

Writing this out can reduce ambiguity.

The Introduction

The introduction should clarify:

Background and Rationale:
Why are we doing this?
Who are we doing this for?
Why is this a reasonable thing to try?

Objectives:
What are we trying to establish?
What outcomes would constitute a “success”?

Data & Methods

Data:

What type of data is on hand?
Where did it come from? Supply URLs if appropriate.
How many samples are being examined?
How many assay measurements per sample?
What covariates are we exploiting?

Methods:

How will we process the data?
Outline or reference the pipeline employed.
What statistical tests will we employ?
How will we adjust for multiple testing?
If the data will be filtered, what cutoffs will be used?

Results

This should contain
an objective summary of our findings;
our interpretation should follow in the Conclusions.

How many samples/genes survived filtration?

How many genes are significant at our chosen FDR?
If there aren't many (e.g., ≤ 10), enumerate them.

Figures and tables in the report can be referenced but do not belong in the summary itself.

Other files produced (e.g., csv files containing p-values and annotations) should be listed by name.

Conclusions

The conclusions should provide context for the results by interpreting results in light of our objectives.

For each objective mentioned in the introduction, state whether that objective has/hasn't been met.

The conclusions may include some discussion of the implications of the findings.

Do the findings make sense? Should we be excited?

The discussion can include data quality issues, caveats or limitations of the approach used, and possible next steps.

Strive for Parallelism

If there are multiple objectives listed in the introduction,
address them in the same order
when discussing the conclusions.

Similarly, if tests A, B, and C are described in the methods,
present the results in the same order.

Parallelism improves clarity and makes it easier to “check off”
whether everything has been done.

What Can We Write Before Analyzing the Data?

Much of the summary can be written before analysis begins.

Indeed, sending your collaborators a draft of the
Introduction, and
Data & Methods
before starting analyses is a *very good idea*.

This prevents wasting time on analyses we don't care about.

Code

I like to precede each block of code with a statement of what I'm trying to achieve.

If the block is intended to process data, I often include small chunks of "before" and "after" data to show how it worked.

Long ($> 1pg$) uninterrupted blocks of code should rarely, if ever, be used.

Write functions for blocks of code that are extensively reused.
Function names should be descriptive.
Use named arguments.

Descriptive Names

Name your variables.

Name your data frames and matrices.
Name their rows and columns.

Try to refer to entries by name, and not by number.

This should clarify your intent.

“Your worst collaborator is yourself, 6 months ago, and you don’t answer emails” – Karl Broman

Appendices

At the end of every report, I include an appendix with calls to

```
> sessionInfo()
```

and

```
> getwd()
```

This helps us find and set things up again.

It also clarified when certain reports hadn't been put on shared drives before being sent out.

Clarity, clarity, clarity

Questions I find myself asking report writers regularly:

Do all team members share an understanding of the goals?

Common changes I request (beyond spelling and grammar):

What do + and - mean?
Is a high value good or bad for the patient?

What do you infer from the plot you're showing?
Describe what lets lets you do this.
What would the plot look like if there's no structure?

What is this chunk of code meant to do?

What Sanity Checks Have You Employed?

What changes did you expect to see? Did you?

What changes shouldn't be there if this is working?

Have you plotted the p-values from all contrasts?

Have you plotted low dimensional summaries of the data? (e.g., PCA?)

How have you plotted the data?

Do the results make sense?

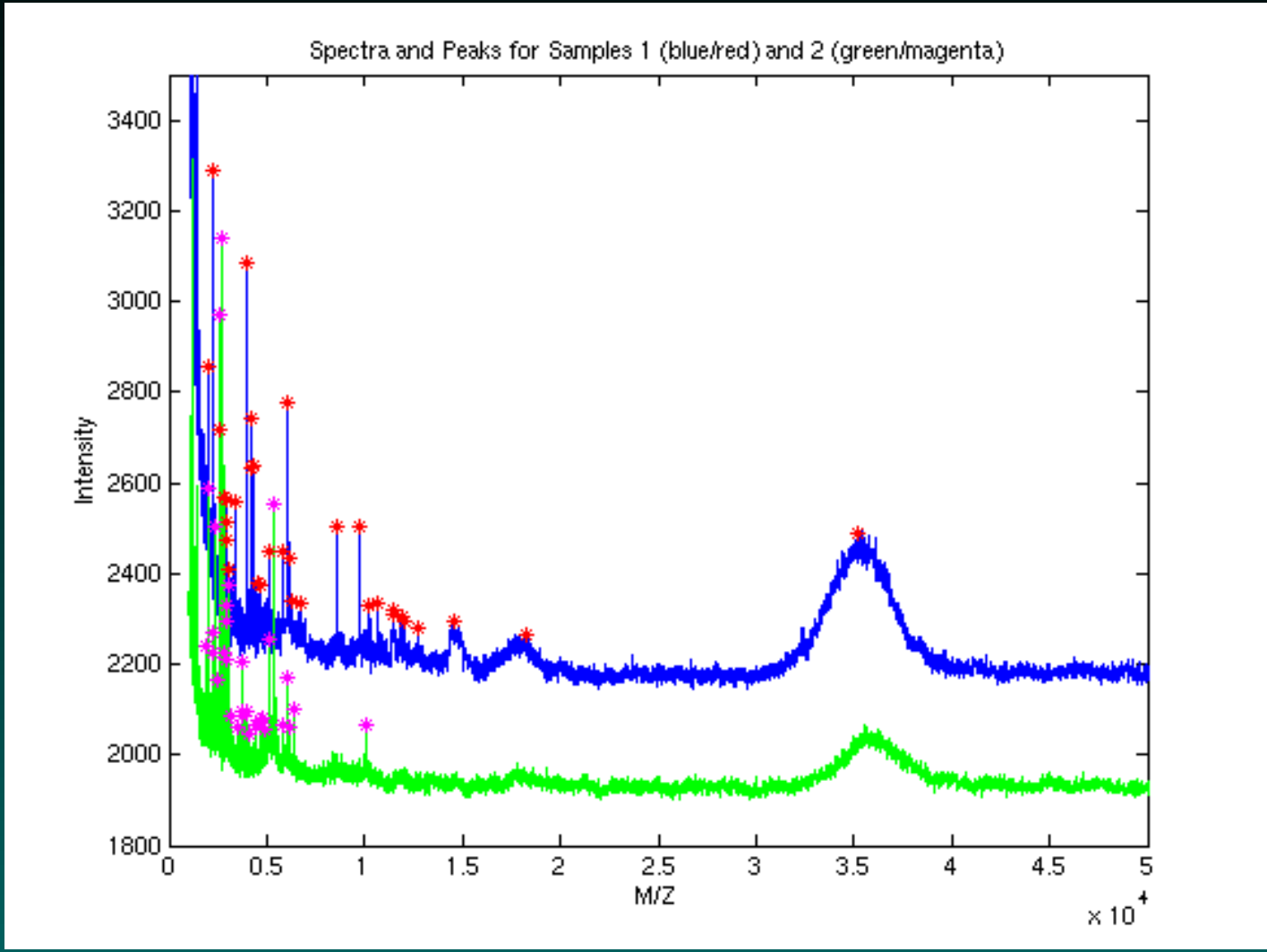
The Proteomics Data Mining Competition

41 samples, 24 with disease*, 17 controls.

20 fractions per sample.

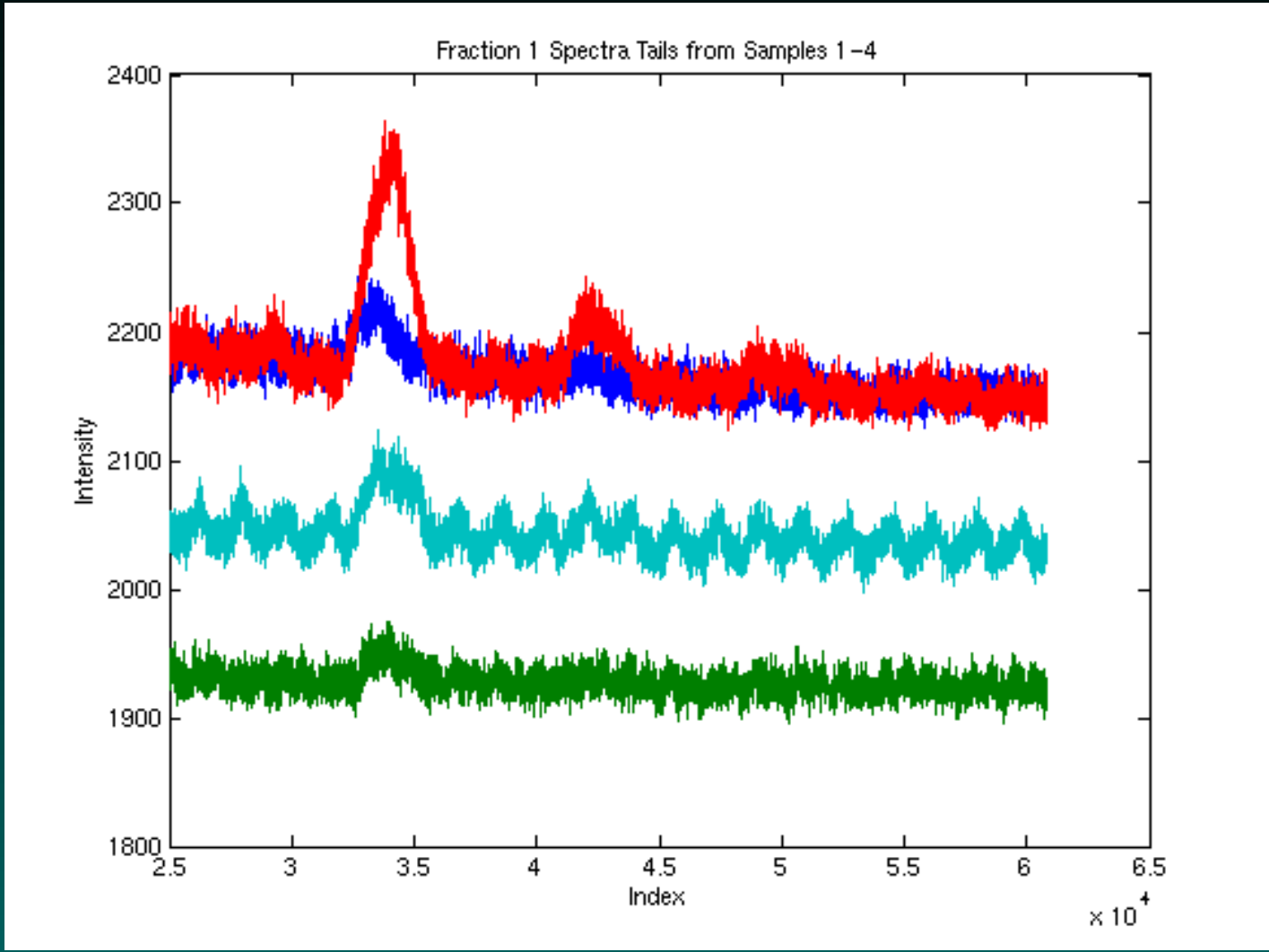
Goal: distinguish the two groups.
We know this can be done due to the “zip effect”.

Raw vs Processed – Use Raw



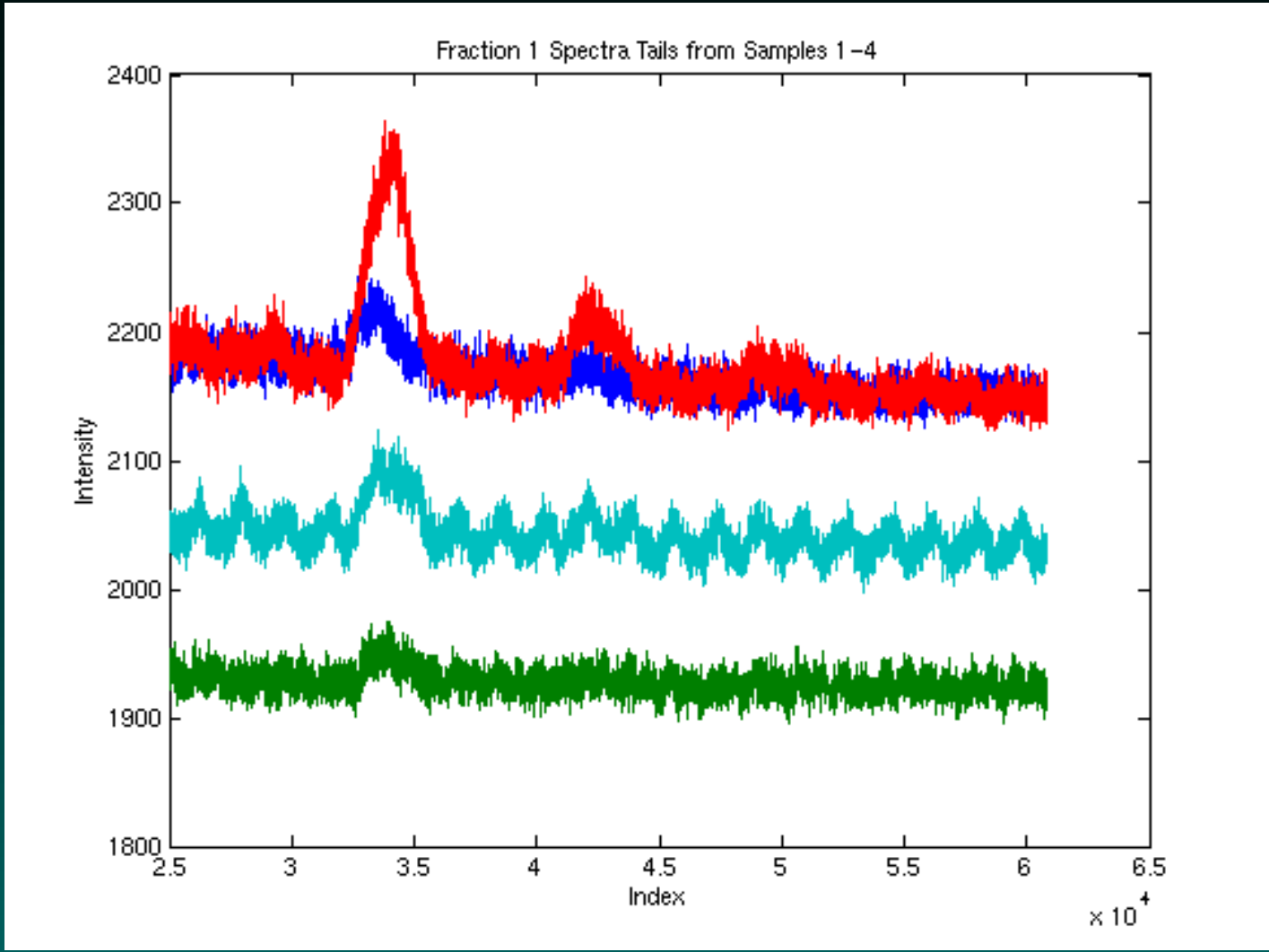
Note the need for baseline correction!

Oscillatory Behavior...



Half the spectra are “wiggly”!

Oscillatory Behavior...



Half the spectra are “wiggly”! It’s the A/C power cord.

The Importance of Communication

If something's worth doing, it's worth doing well.

Teams work better if everyone has some common understanding of what's going on.

Part of this is "making a pitch"

Why should people care?

Before anyone will attempt reproduction, they need to be persuaded that it matters.

(Lessons from Media Relations and 60 Minutes)