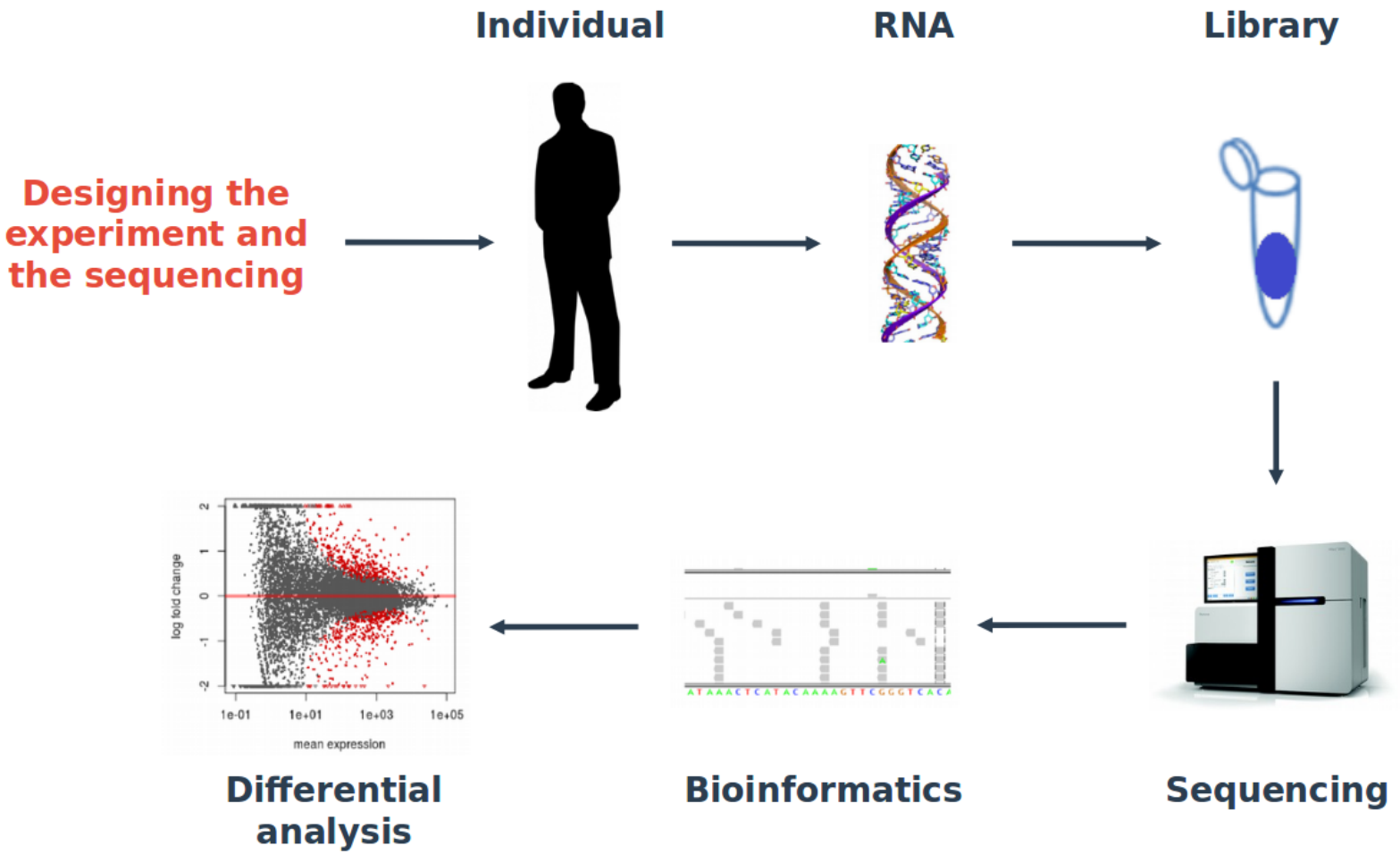
Keep control on your data

Thibault Dayris, Daniel Gautheret

2018-01-08

Introduction

Main RNA-Seq steps



Advices

"To consult a statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination. He can *perhaps* say what the experiment died of."
Ronald A. Fisher, Indian Statistical Congress, 1938, vol. 4, p 17

Advices

"While a good design does not guarantee a successful experiment, a suitably bad design guarantees a failed experiment"

Kathleen Kerr, Atelier Inserm 145, 2003

Vocabulary

Sample	Variable	Factor
Replicate A-1	Level A	Biological condition X
Replicate A-2	Level A	Biological condition Y
Replicate A-3	Level A	Biological condition Z
Replicate B-1	Level B	Biological condition X
Replicate B-2	Level B	Biological condition Y
Replicate B-3	Level B	Biological condition Z

Statistical Modeling

Goal: address **one** biological question.

Statistical modeling consists in using mathematical formulas involving:

- Experimental conditions X
- Numerical values measured Y
- Parameters β linking X and Y (to be estimated), *e.g.*: $Y \sim X\beta + \varepsilon$
- Some hypotheses on the data variability, *e.g.*: $\varepsilon \sim \mathit{Gaussian}(0, \sigma^2)$

Experimental Design

Goal

To **keep control over the variability** during the experiment, you have to know:

- What is the biological question ?
- How to estimate the associated biological variability ?
- How to control the technical variability ?

Biological or technical uncontrolled effects could:

- Hide/Cancel the biological effect of interest
- Wrongly increase the biological effect of interest

Basic comparison

We are interested in the transcriptome of Tumor and Normal tissues from patients:

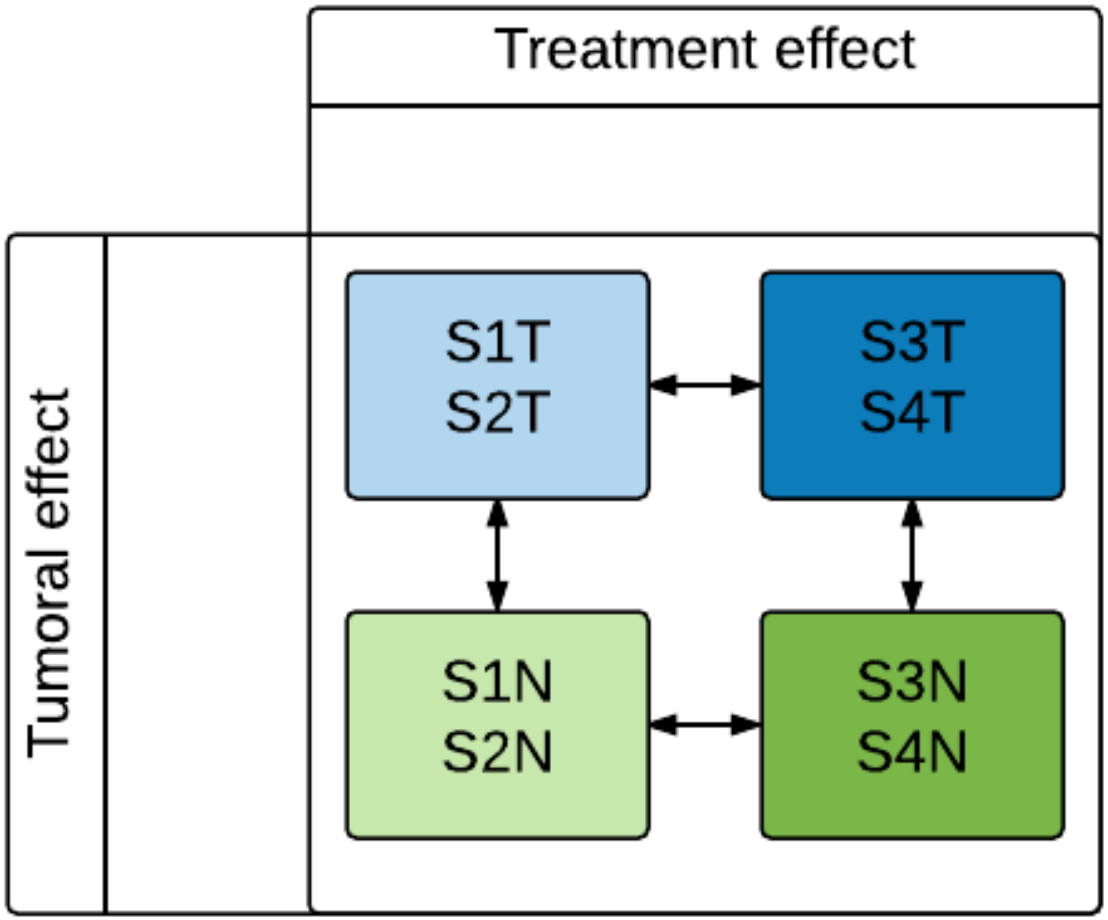
Sample	State
S1T	Tumor
S2T	Tumor
S3T	Tumor
S4T	Tumor
S1N	Normal
S2N	Normal
S3N	Normal
S4N	Normal

Paired Samples

We can add information:

Sample	State	Treatment
S1T	Tumor	Drug 1
S2T	Tumor	Drug 1
S3T	Tumor	Drug 2
S4T	Tumor	Drug 2
S1N	Normal	Drug 1
S2N	Normal	Drug 1
S3N	Normal	Drug 2
S4N	Normal	Drug 2

Interactions between factors and variables



Confounding effects

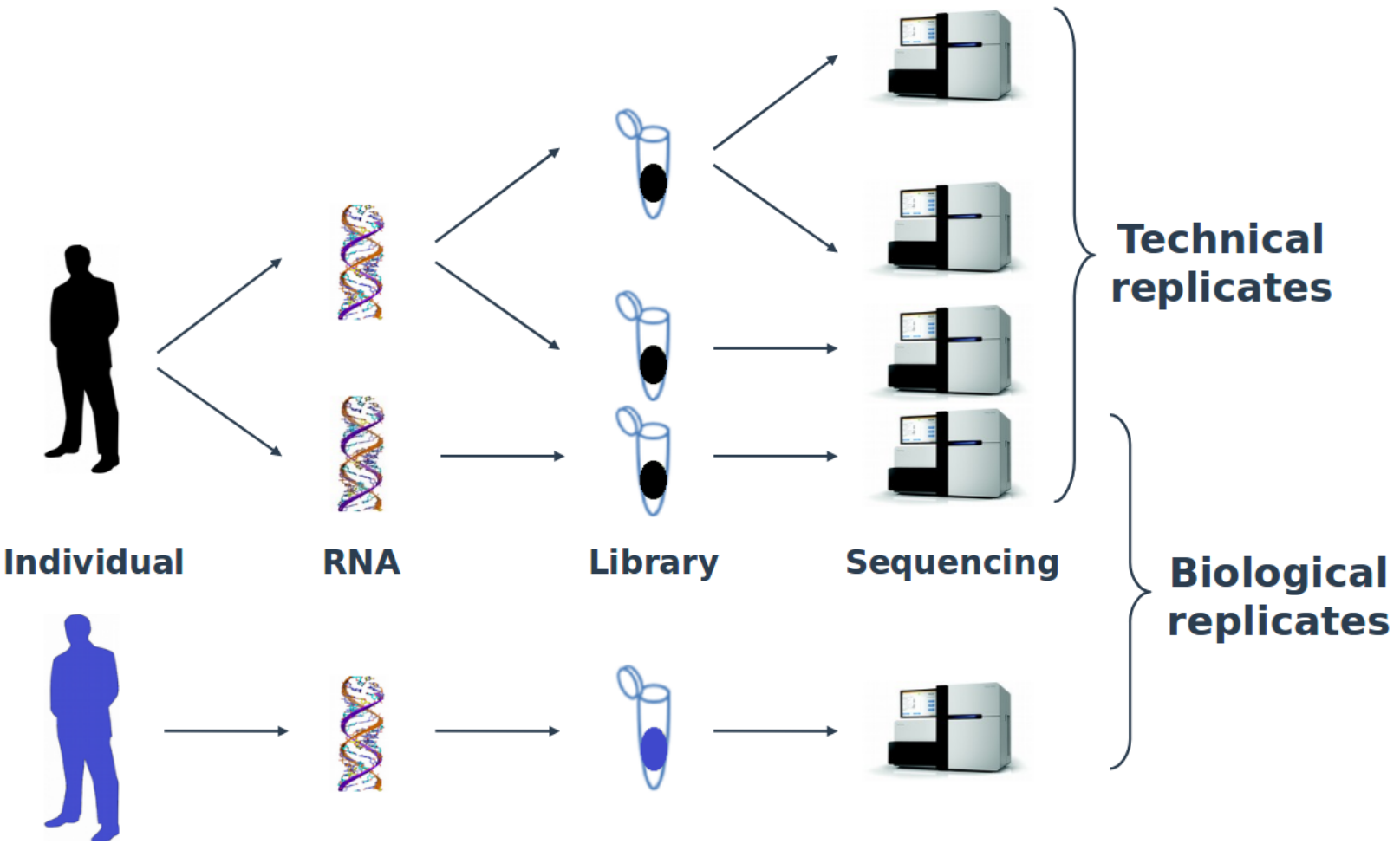
Sample	State	Age	Gender	Extraction date	Experimentalist
H1	Healthy	24	M	02/16/2006	John
H2	Healthy	29	M	02/10/2006	John
T1	Tumor	55	F	07/01/2015	Olivia
T2	Tumor	60	F	07/07/2015	Olivia

Consequences of confounding effects

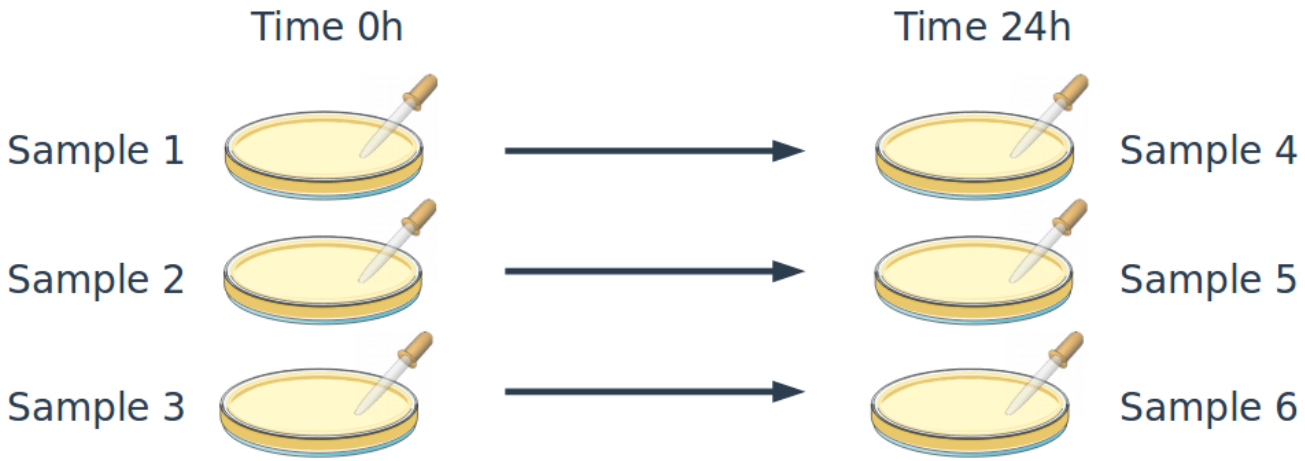
A gene is detected as being differentially expressed between healthy and tumor patients. Is it due to:

- The disease ?
- The age effect ?
- The gender effect ?
- The date ?
- The technician effect ?

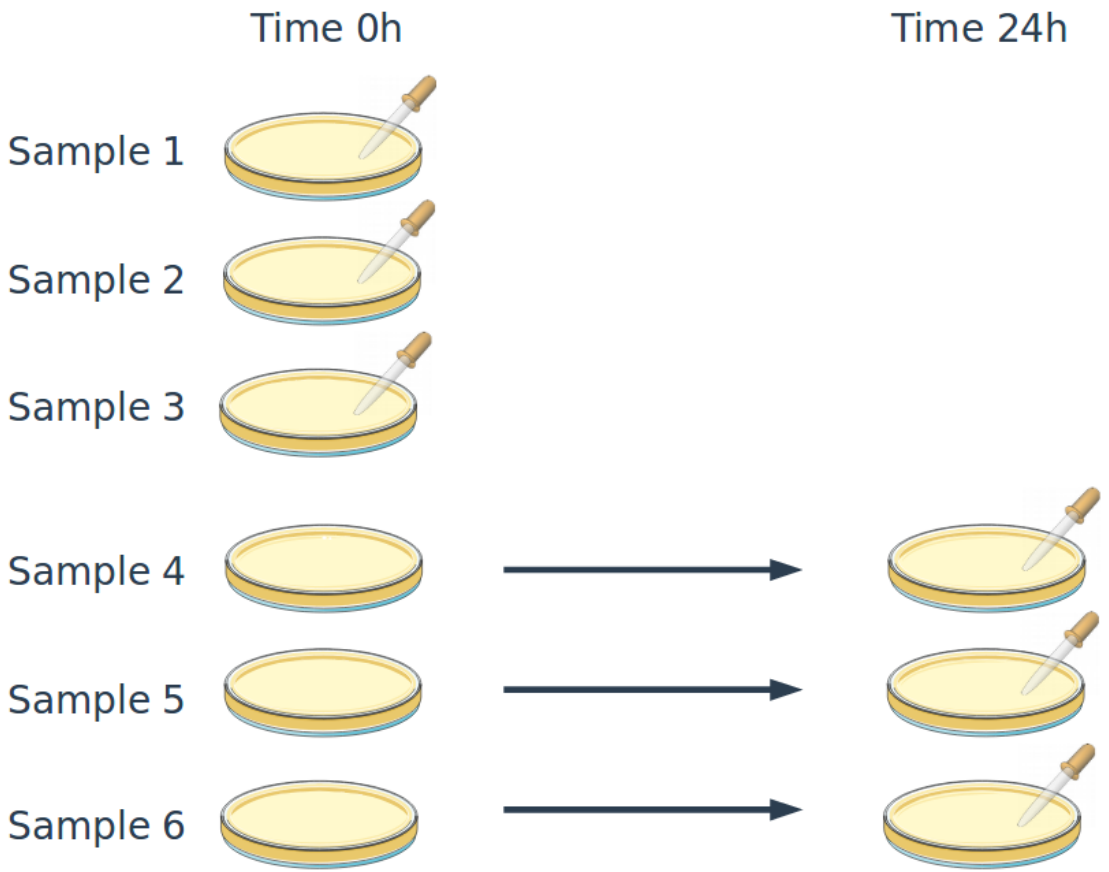
Biological vs. technical replicates



Example of cell lines



Example of cell lines



Importance of the number of biological replicates

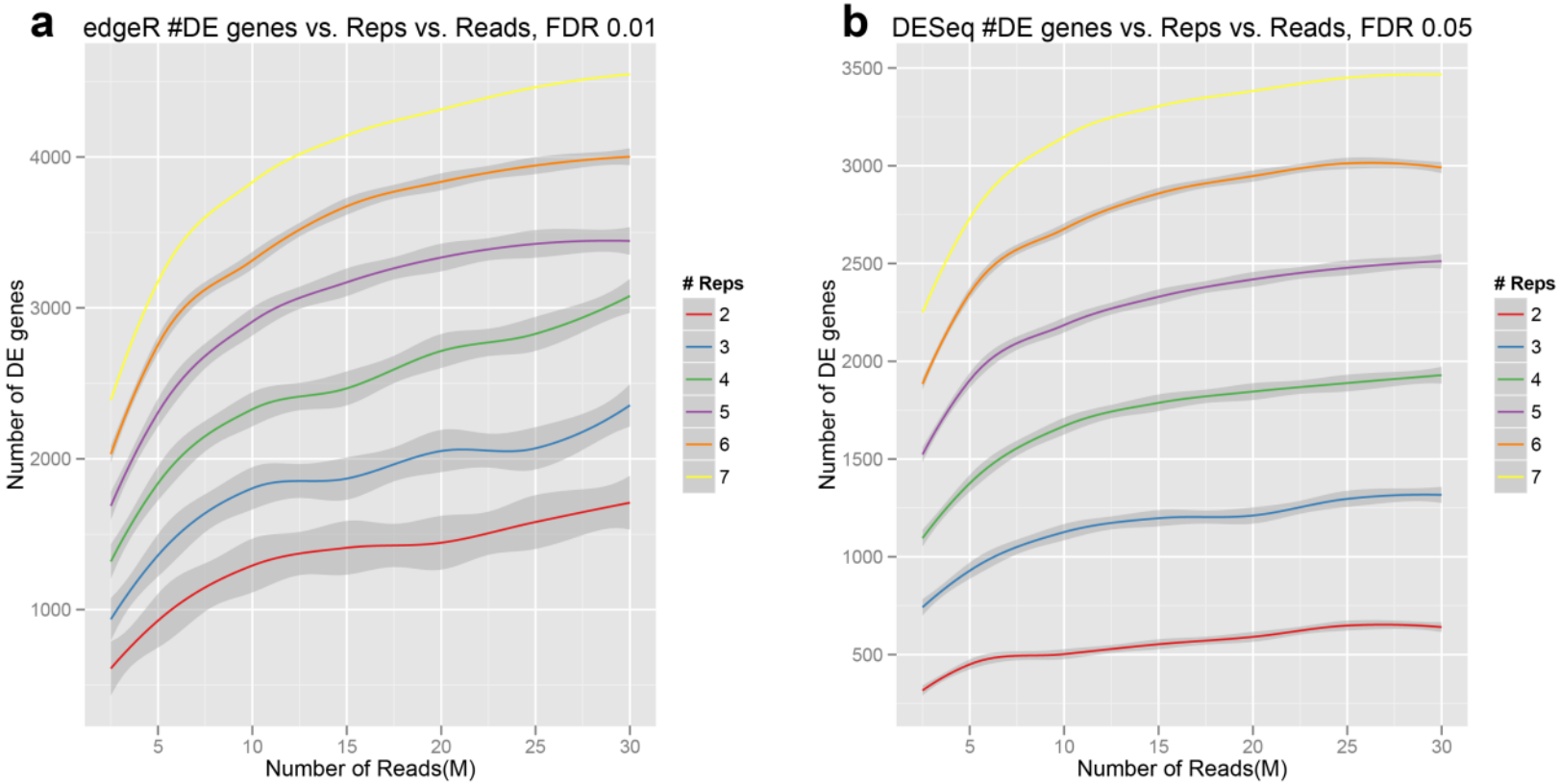
Due to high cost of RSA-Seq, you may want to have 2 or 3 replicates.

With more biological replicates:

- Better estimation of:
- the variability present in the population studied
- the difference between biological conditions
- Better control over the FDR[1]
- Higher statistical power (Differentially expressed targets are more easily detected)

[1] C. Soneson and M. Delorenzi. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics, 14, 2013.

More sequence or more replication ?



Biological replicates increase the number of DE genes identified[2]

[2] Liu, Y., Zhou, J., & White, K. P. (2013). RNA-seq differential expression studies: more sequence or more replication?. Bioinformatics, 30(3), 301-304.

Conclusion on experimental design

You have to **think about your question before** starting to prepare your samples.

At least 3 biological replicates are needed for a differential expression analysis. 80% ~ 60% of your information is **lost** under 4 replicates. Good results are aquired at 6 biological replicates, 12 biological replicates where wide or rare events are searched.[3]

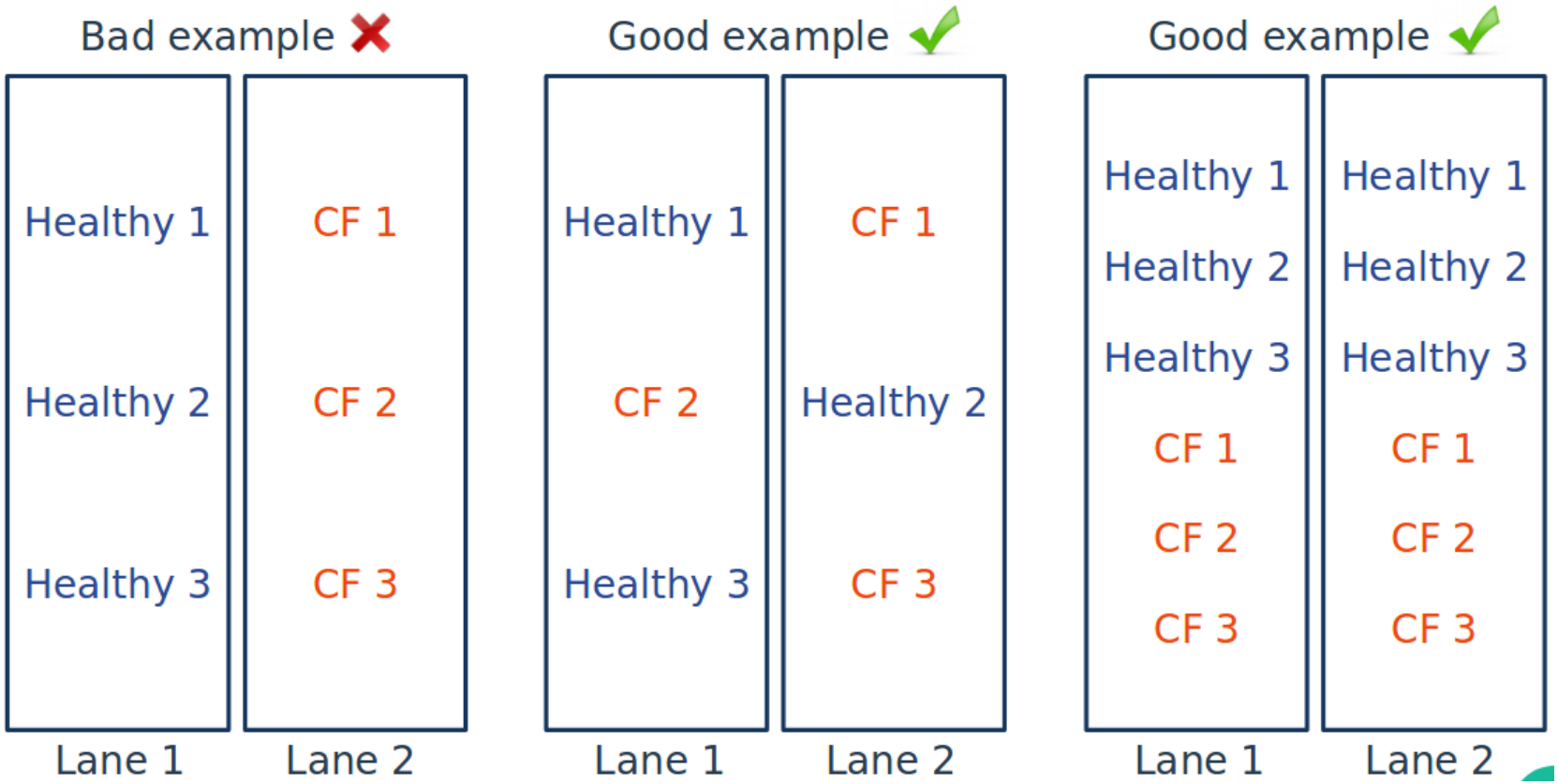
Do not hesitate to ask a bioinformatician/biostatistician in upstream work phase.

[3] Schurch, N. J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., ... & Blaxter, M. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?. Rna, 22(6), 839-851.

Sequencing design

Lane

Do not add confounding technical effect:



Sequencing design effect

Technical variability includes:

- Lane
- Flow cell
- Run

Usually, we observe:

lane effect $<$ flow cell effect $<$ run effect $<<$ biological variability

Deep sequencing

If you are looking for rare effects (SNP, SNV, Fusions, tight differential expression), then choose a deeper sequencing protocol:

- 15M reads for differential **gene** expression
- 30M reads for differential **transcript** expression
- 80M reads for fusions search
- 100M reads for rare events

If you are not interested in very rare effects, raise the replicate number. Thus, the sequencing will be cheaper !

Single end vs. Paired end

Single-end sequencing perfectly fits the following goals:

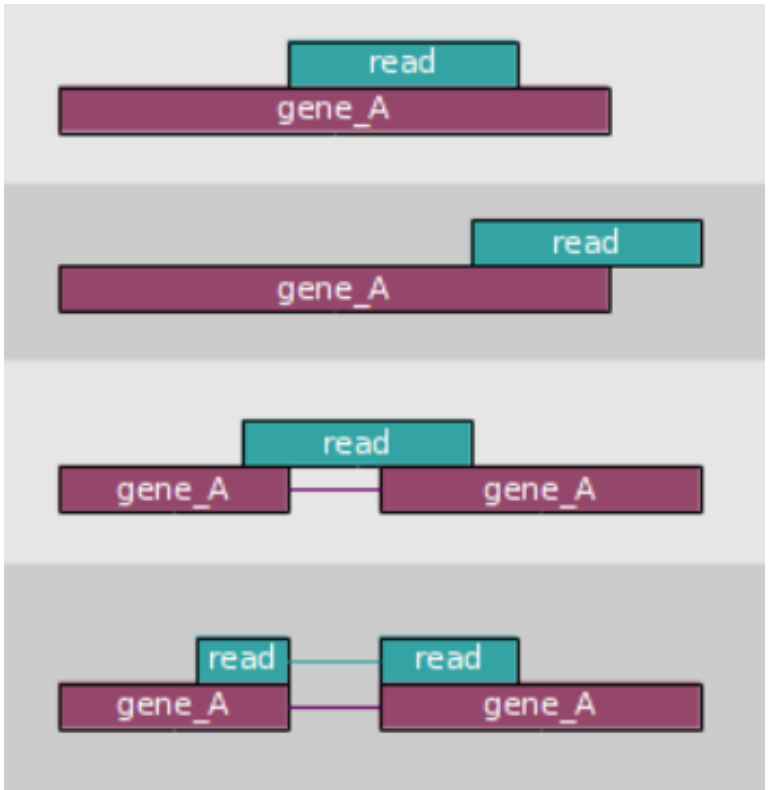
- Gene expression
- Well known genome
- Global expression overview

Paired-end sequencing perfectly fits the following goals:

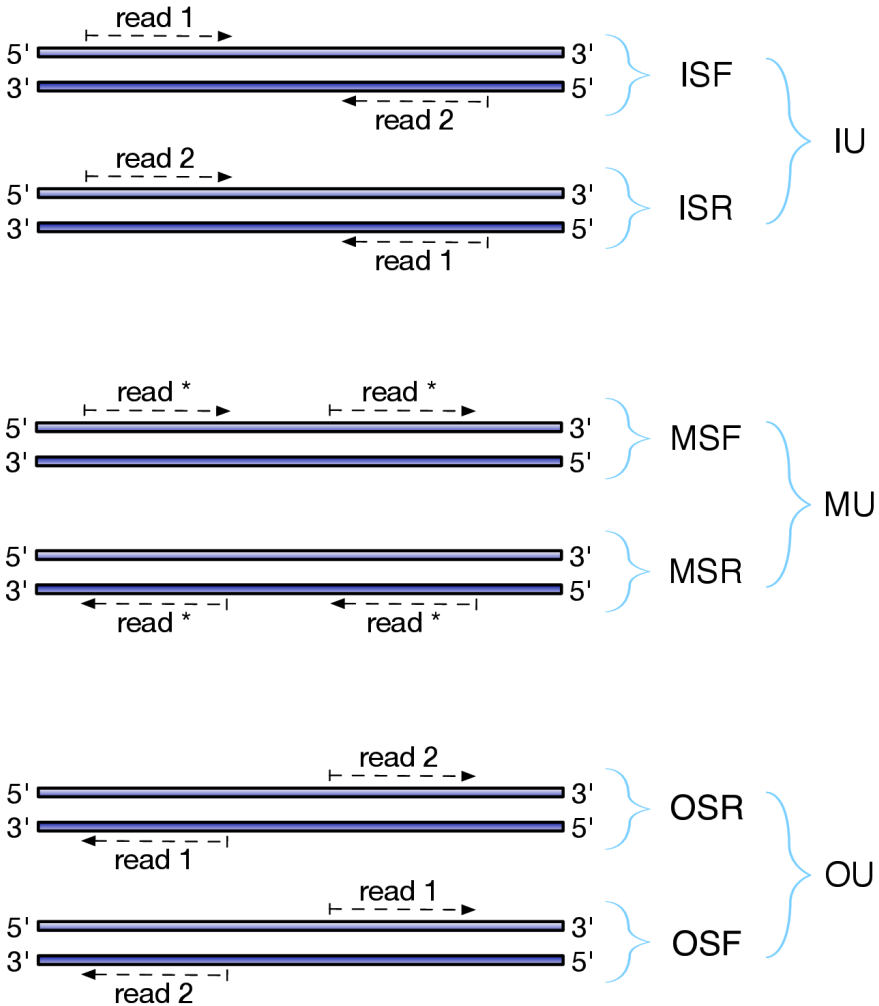
- Transcript expression
- Rare events search
- New isoform discovery
- Fusion / Translocation search

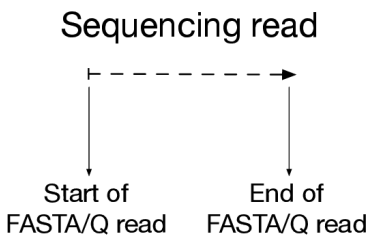
Remember the adage: "Who wants most, can do least"

Library orientation



Library orientation





Conclusion on sequencing design

Talk to your sequencing platform. They always can give you sequencing reports with multiple information:

Sequencer model

- Hiseq, SOLid, Ion ...

Conclusion on sequencing design

Talk to your sequencing platform. They always can give you sequencing reports with multiple information:

Sequencer model

- Hiseq, SOLid, Ion ...
- Quality score encoding

Conclusion on sequencing design

Talk to your sequencing platform. They always can give you sequencing reports with multiple information:

Sequencer model

- Hiseq, SOLid, Ion ...
- Quality score encoding
- Choice of downstream bioinformatics tools

Conclusion on sequencing design

Talk to your sequencing platform. They always can give you sequencing reports with multiple information:

Sequencing kit

- Adapter list

Conclusion on sequencing design

Talk to your sequencing platform. They always can give you sequencing reports with multiple information:

Sequencing kit

- Adapter list
- Expected read size

Conclusion on sequencing design

Talk to your sequencing platform. They always can give you sequencing reports with multiple information:

Sequencing kit

- Adapter list
- Expected read size
- Expected insert size

Conclusion on sequencing design

Talk to your sequencing platform. They always can give you sequencing reports with multiple information:

Sequencing kit

- Adapter list
- Expected read size
- Expected insert size
- Paired protocol ?

Conclusion on sequencing design

Talk to your sequencing platform. They always can give you sequencing reports with multiple information:

Sequencing kit

- Adapter list
- Expected read size
- Expected insert size
- Paired protocol ?
- Library orientation ?

Conclusion on sequencing design

Talk to your sequencing platform. They always can give you sequencing reports with multiple information:

Sequencing kit

- Adapter list
- Expected read size
- Expected insert size
- Paired protocol ?
- Library orientation ?
- Expected content of your sequencing results ?

Conclusion on sequencing design

Talk to your sequencing platform. They always can give you sequencing reports with multiple information:

Biological sample quality:

- RIN, rRNA ratio, ...

Conclusion on sequencing design

Talk to your sequencing platform. They always can give you sequencing reports with multiple information:

Biological sample quality:

- RIN, rRNA ratio, ...
- Expected bias ?

Quality control

FastQC, a tool designed to quality control

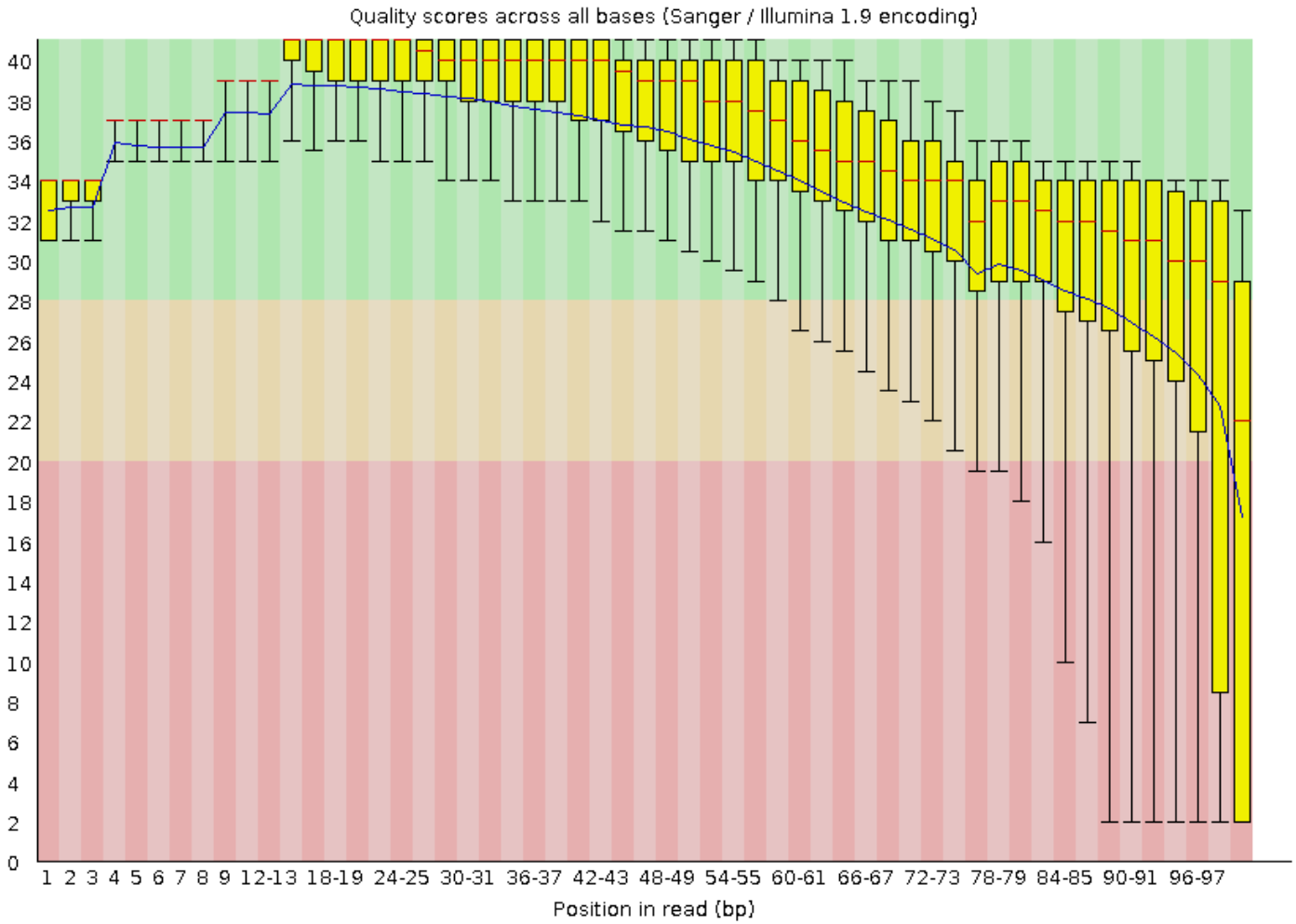
The sequencing platform may give you your results under the following formats:

- fastq: list of reads and their quality
- bam: list of reads mapped against a genome
- ubam: list of unmapped reads written in bam format

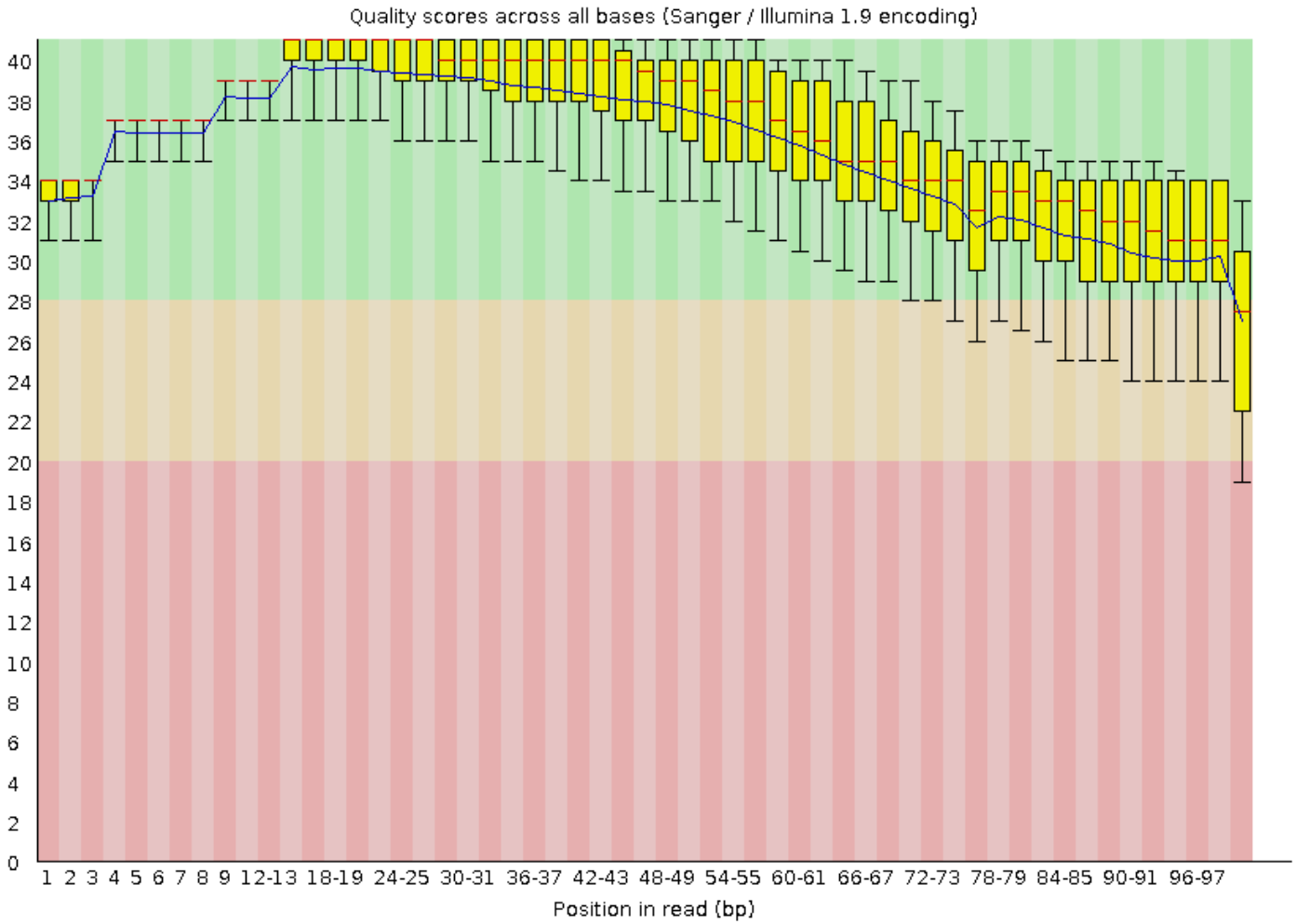
Keep control over your data: check them with FastQC[4]

[4] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

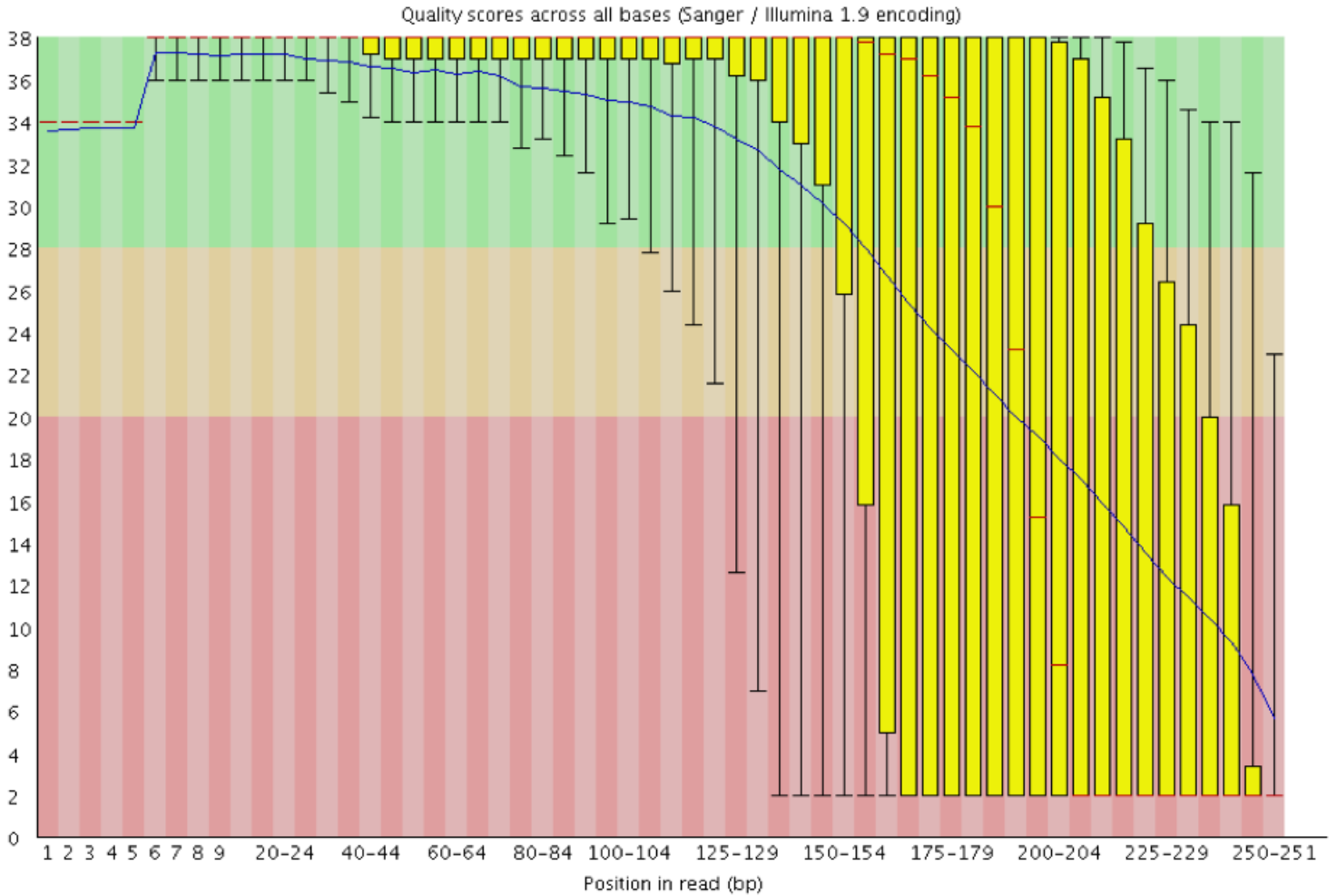
Per base quality score (untrimmed data)



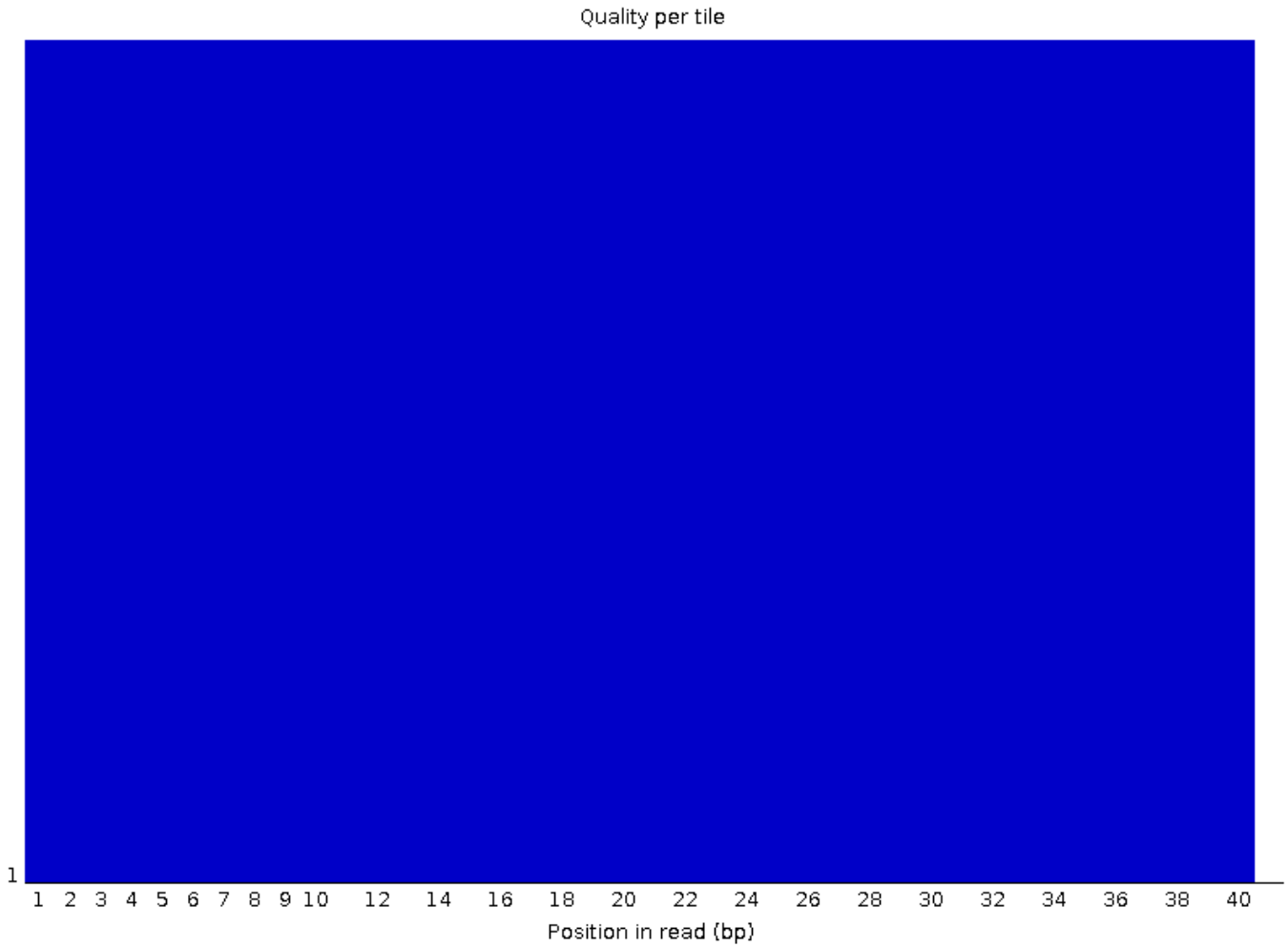
Per base quality score (trimmed data)



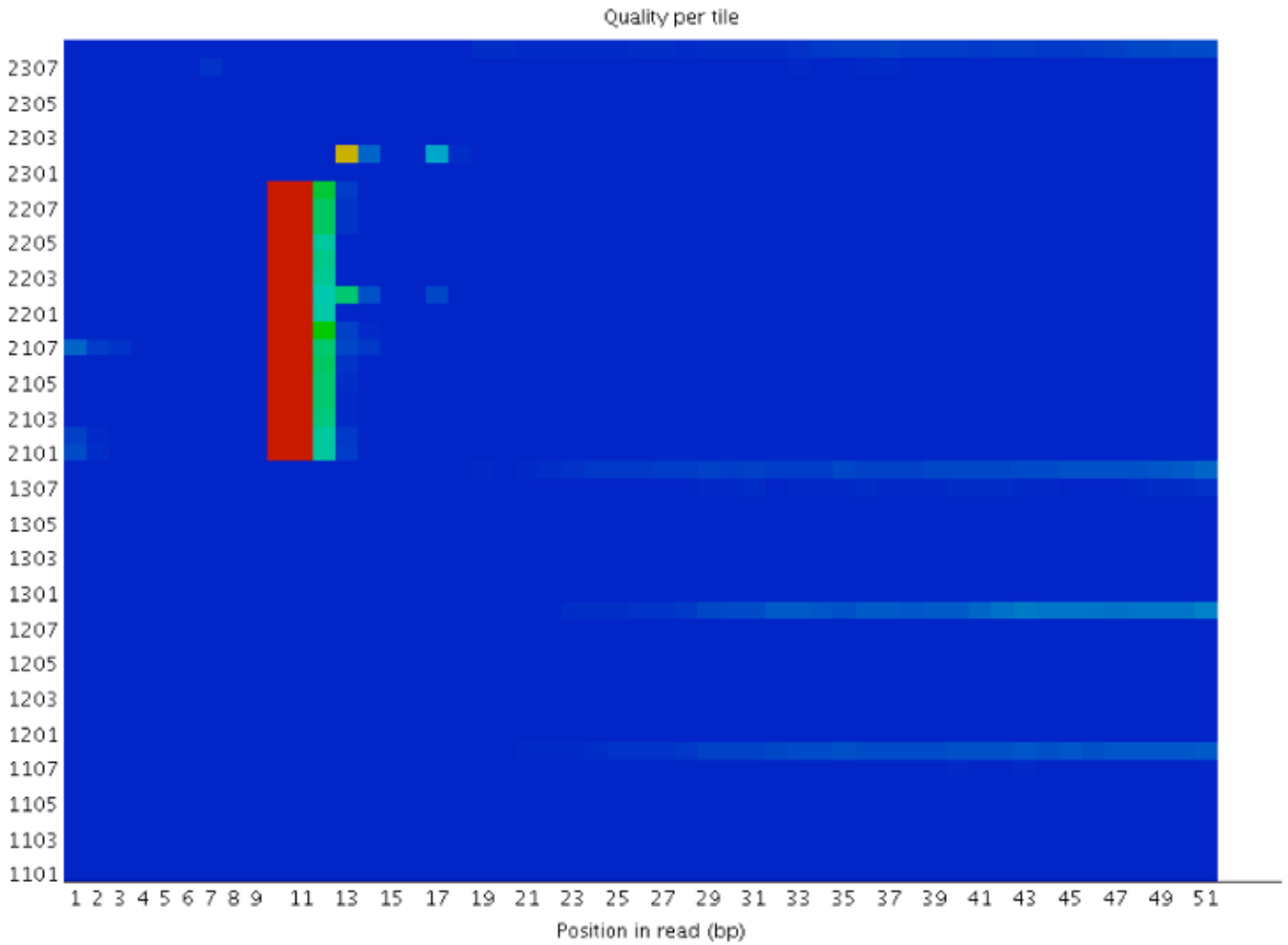
Classic RNA-Seq quality loss



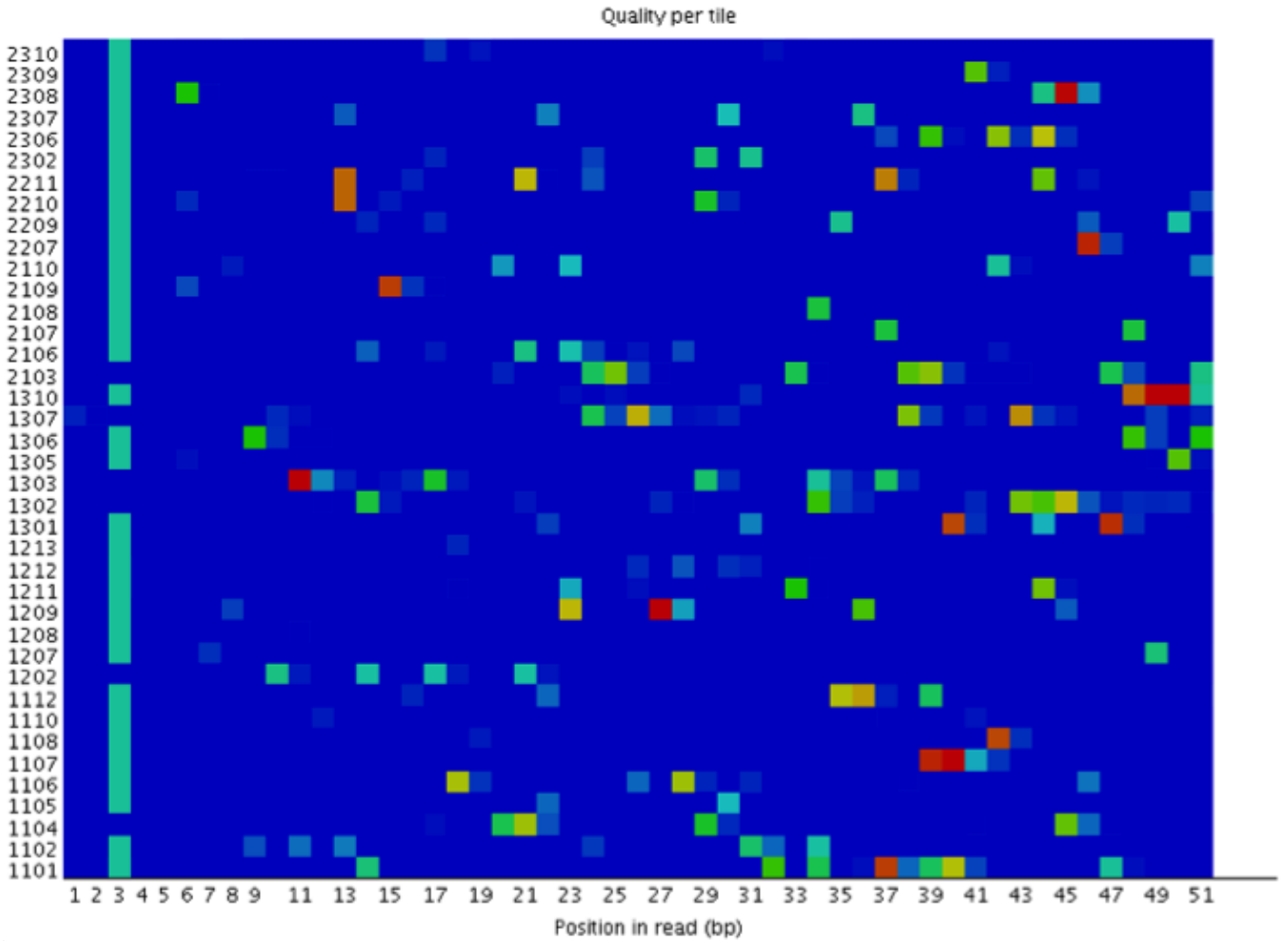
Per tile quality score



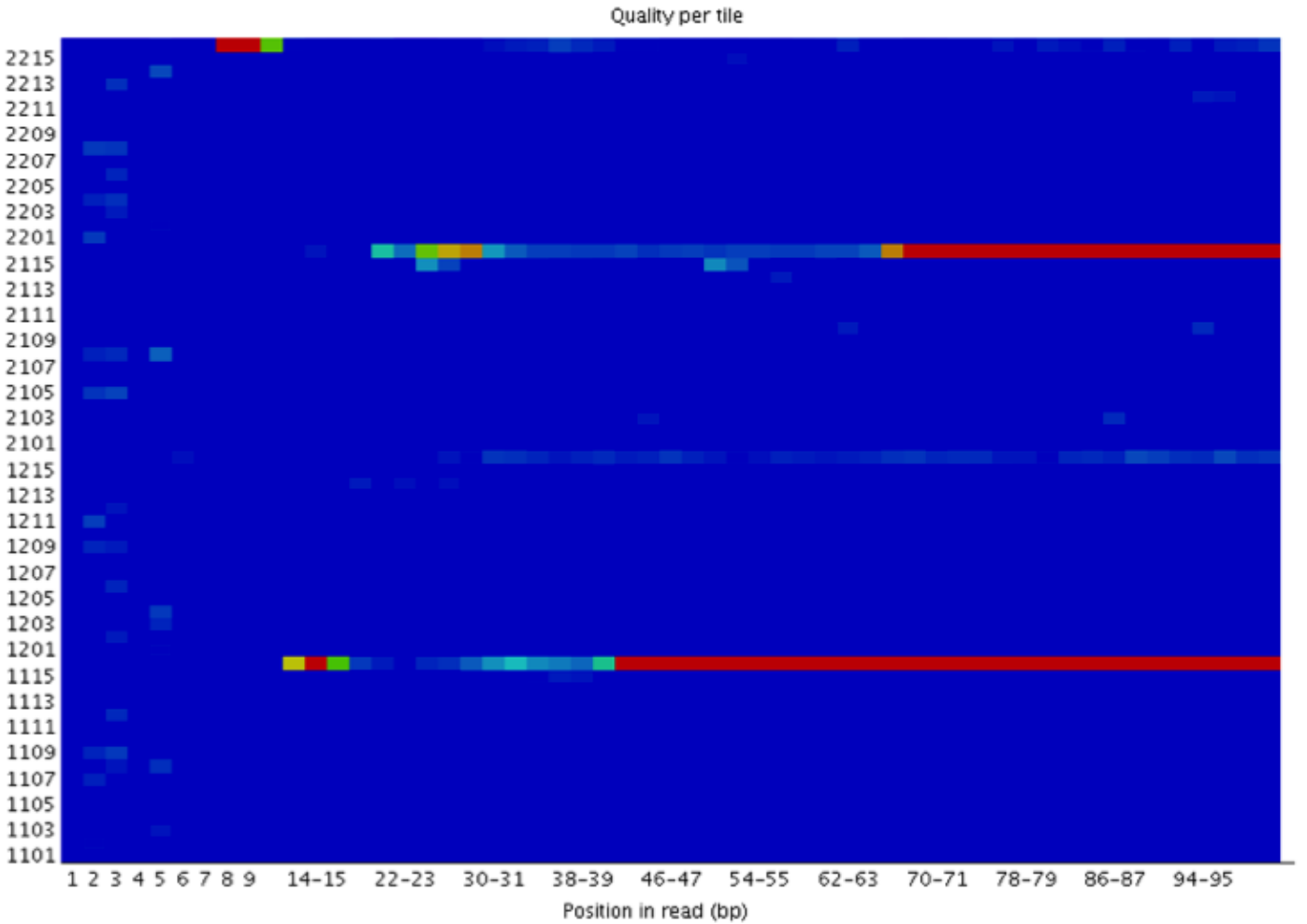
Per tile quality score: bubble ?



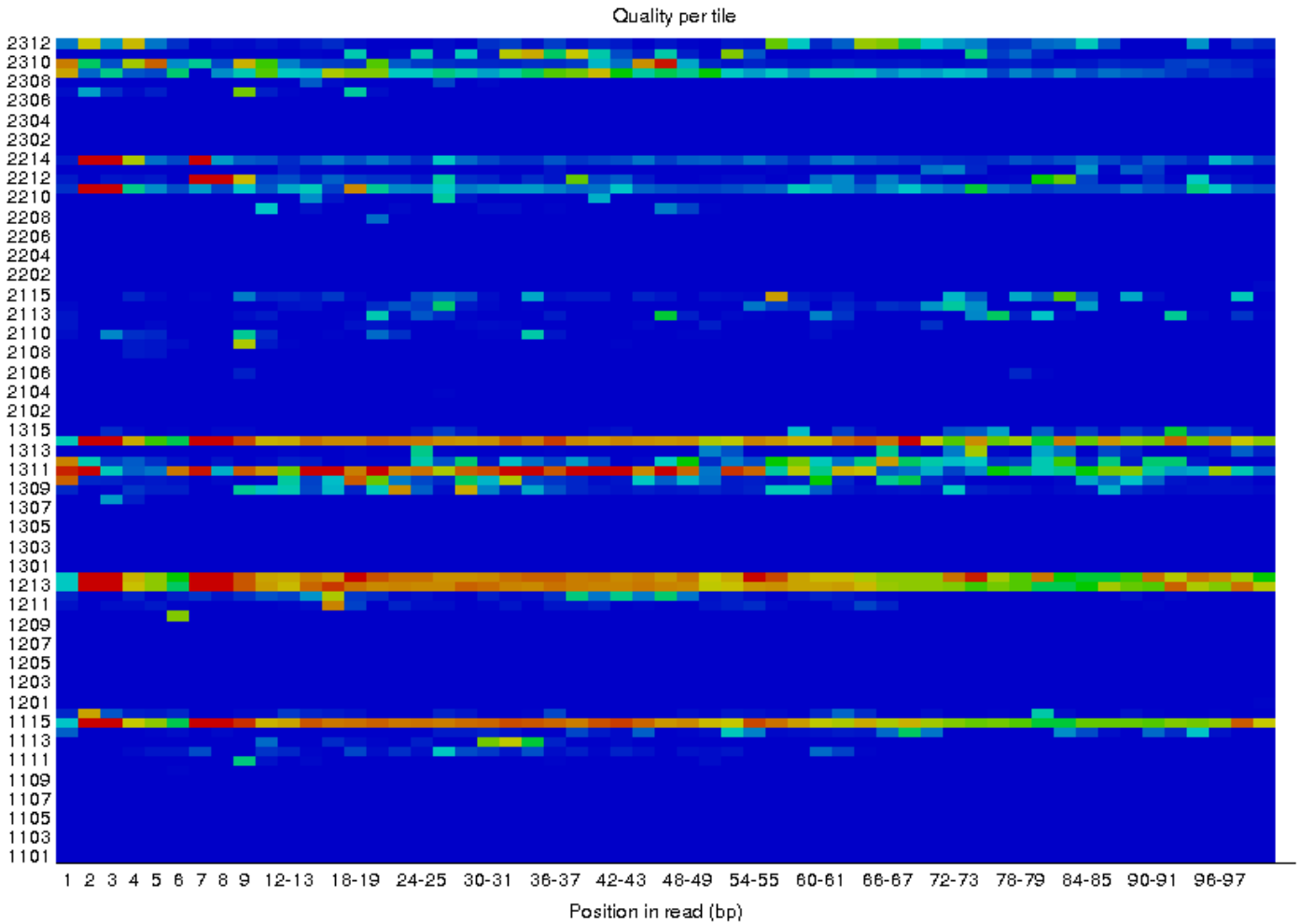
Per tile quality score: overloading ?



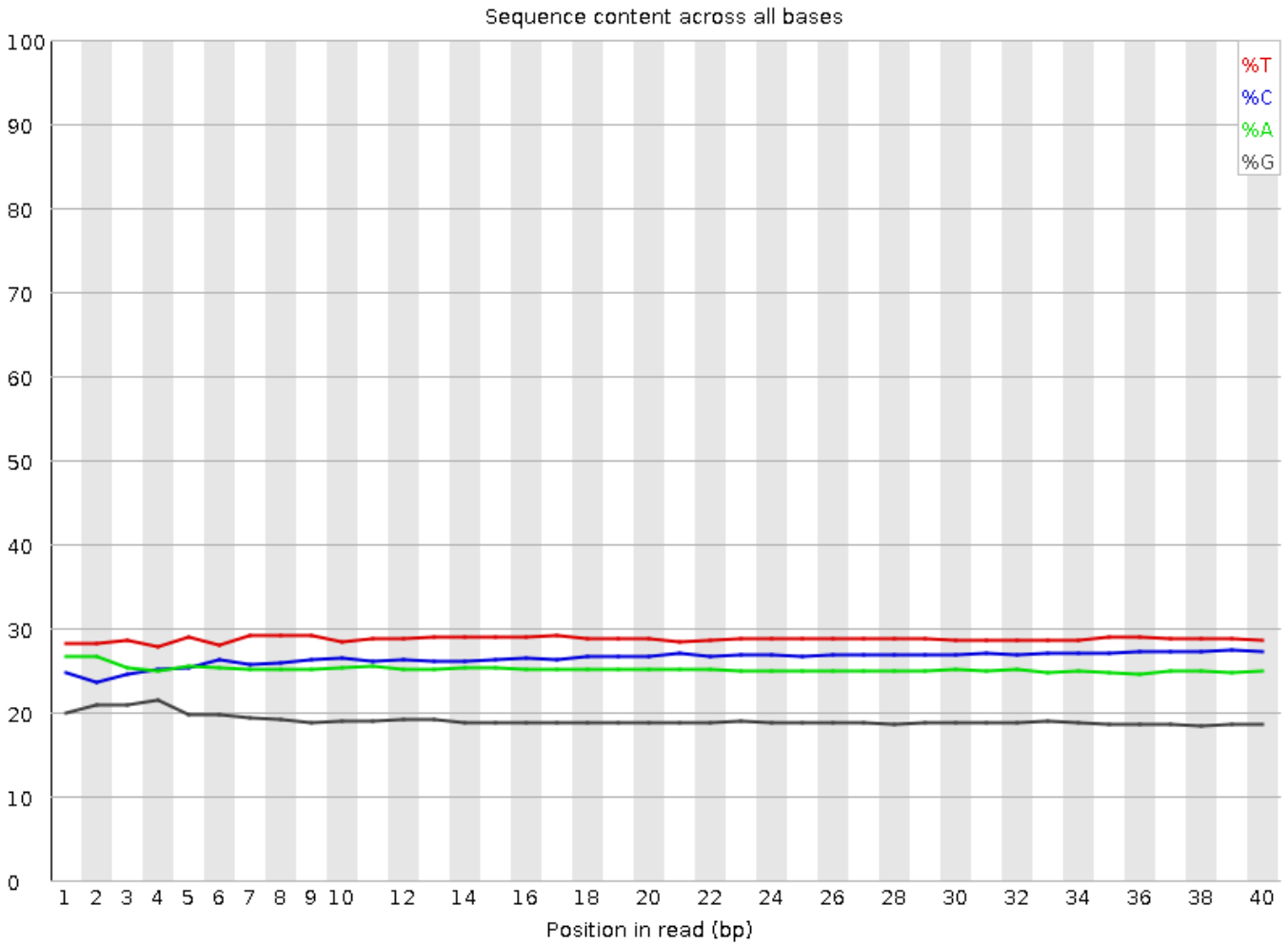
Per tile quality score: Obstruction ?



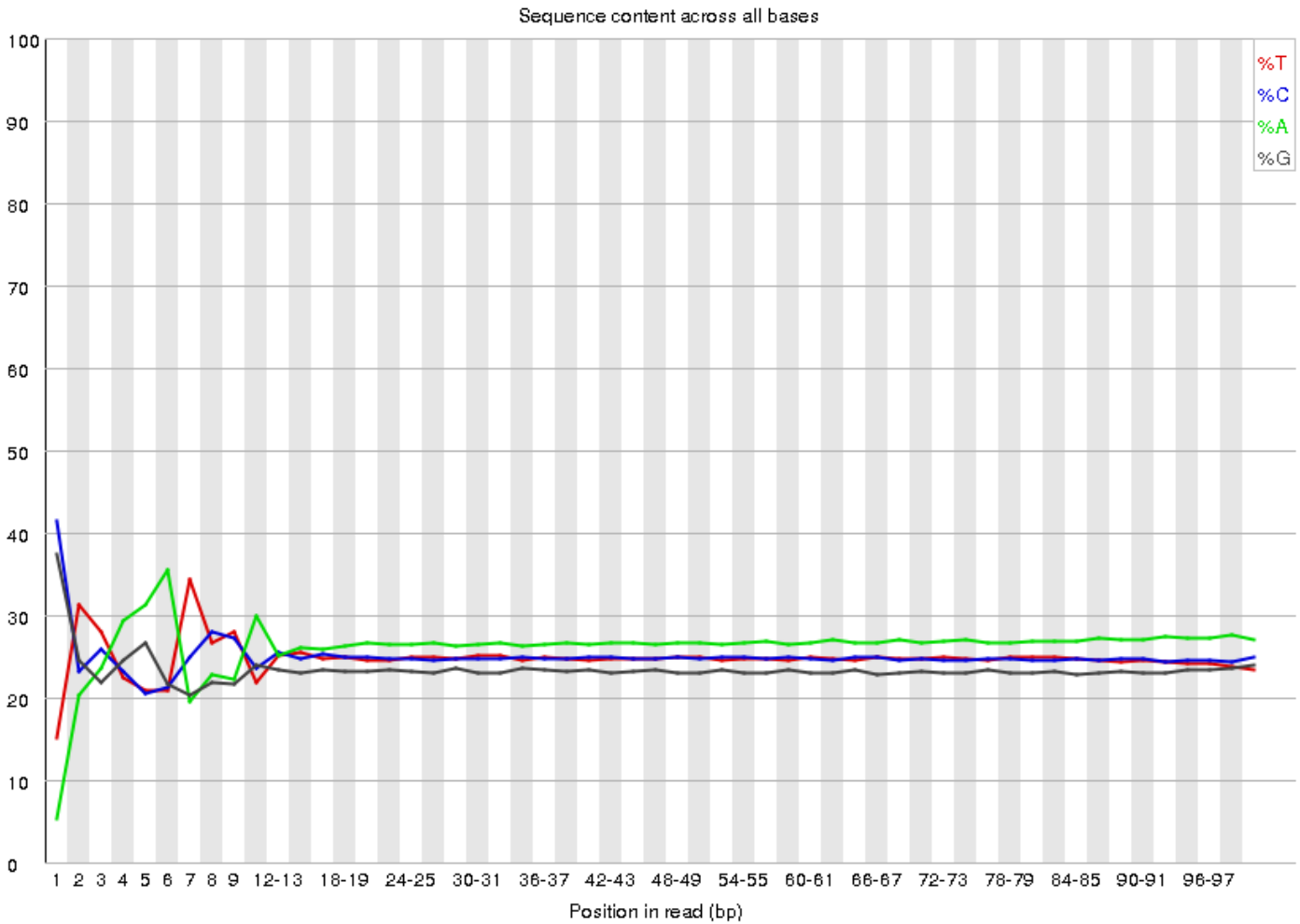
Per tile quality score: Poor tile



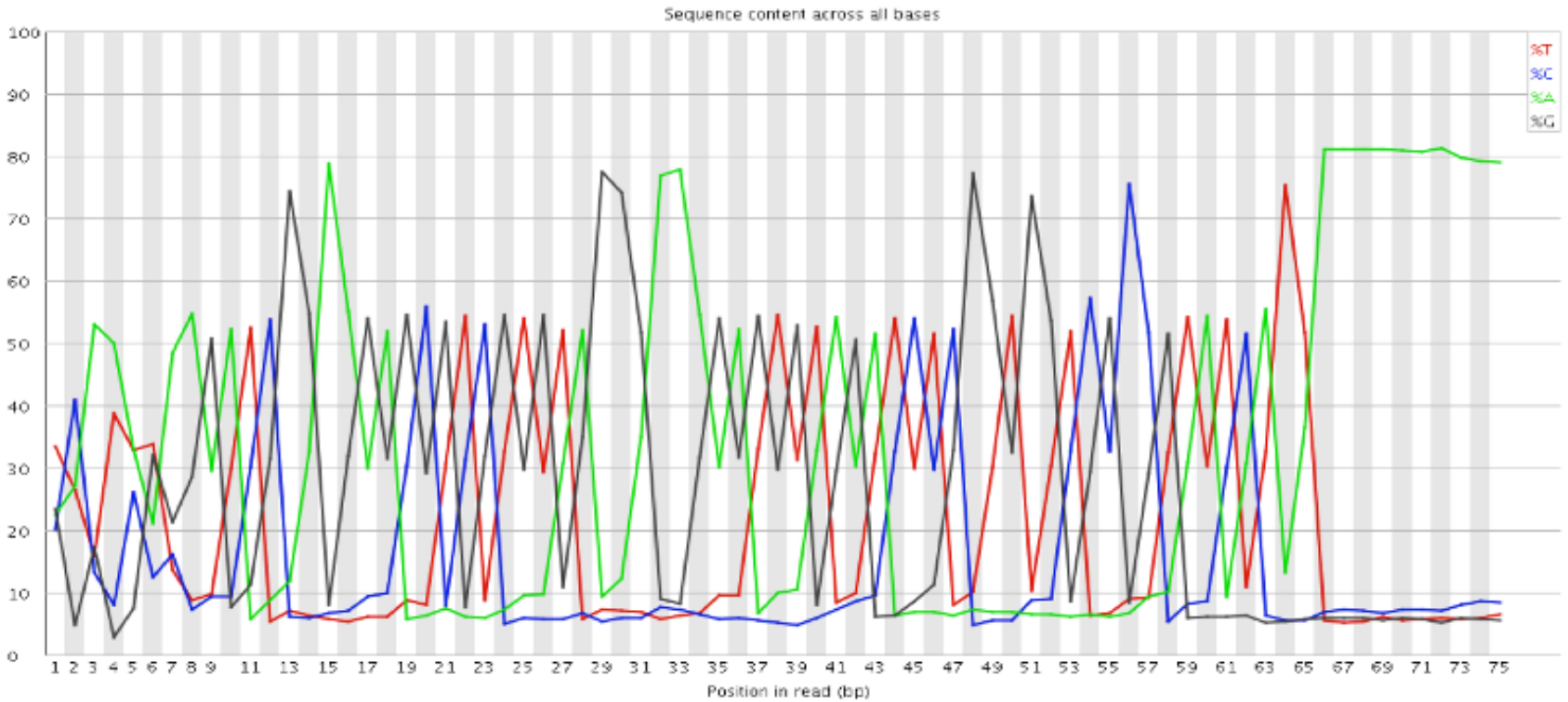
Sequence content



Sequence content: Tipycal bias



Sequence content: classic contamination



Quality control conclusions

FastQC contains many other graphs which are very informative (repeated sequences, contamination, . . .)

Remember that most of the warnings raised by this tool are explained by expected variations in your samples (IG enrichment, poly-A selection, PCR cycles, . . .)

Your knowledge of your data (spices, disease, treatment, . . .) will help your bioinformatician team mate

Conclusions

General conclusions

A RNA-Seq project **requires discussions** between biologists, bioinformaticians, and biostatisticians **as soon as the project starts**!

Statistics and bioinformatics can not give you "everything" on your data set. **One question, one experiment.**

Thanks

Special thanks to Marc Deloger[5], and Hugo Varet[6] who helped me to build this presentation.

Thank you for your attention

[5] Institut Curie
[6] Institut Pasteur