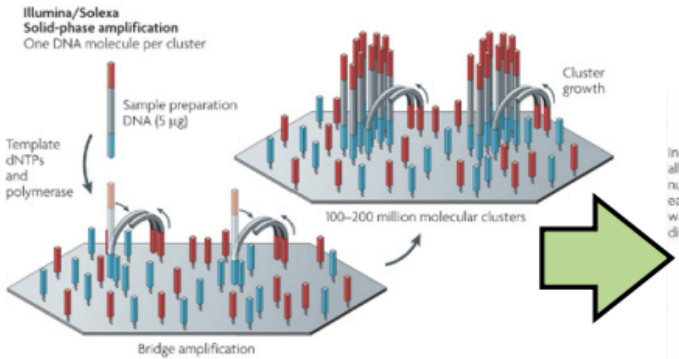
Pre-processing and normalization

Jeff Leek

@jtleek

www.jtleek.com

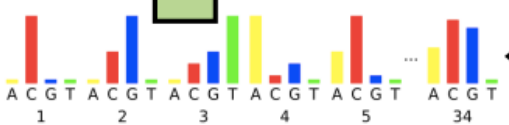
http://www.cbcb.umd.edu/~hcorrada/CMSC858B/lectures/lect22_seqIntro/seqIntro.pdf



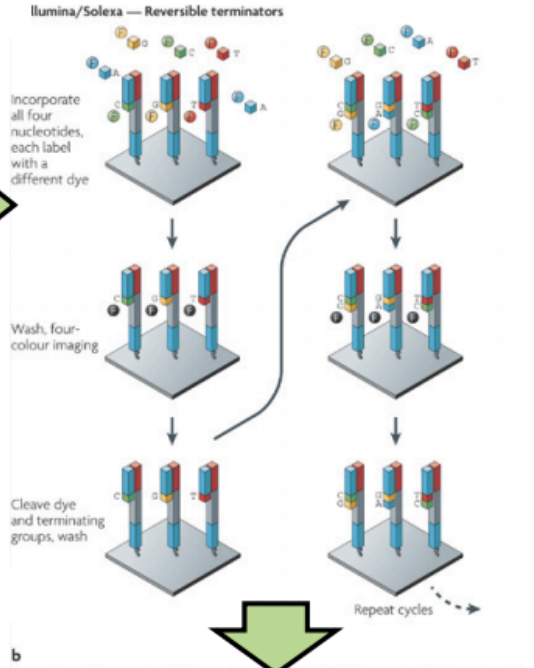
Source: Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010

name
sequence
quality scores

x 100s of
millions



Source: Whiteford et al. Swift: primary data analysis for the Illumina Solexa sequencing platform. Bioinformatics. 2009



Source: Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010

Preprocessing

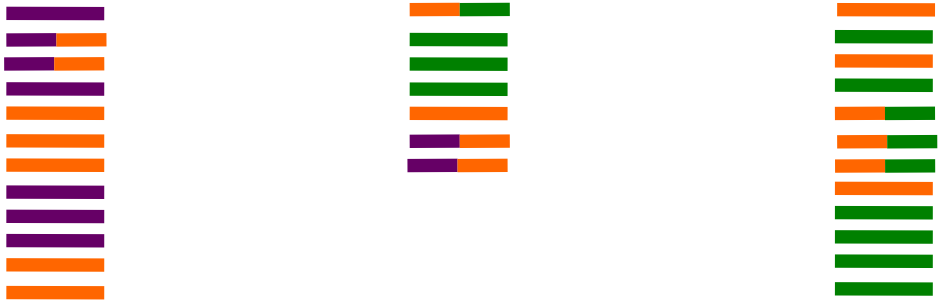
- Convert raw data to “processed”
- Try to remove technological artifacts

Slide courtesy: Alyssa Frazee

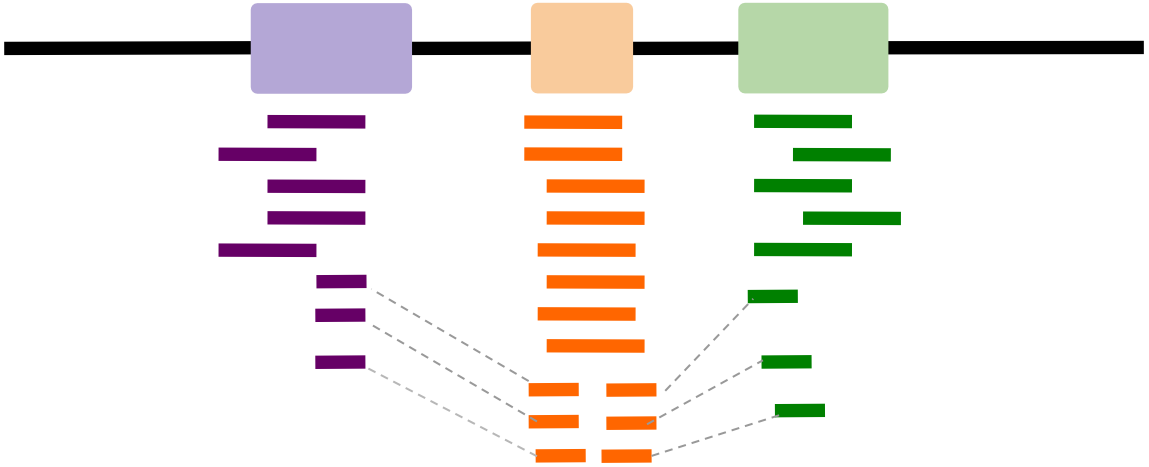
Genome (DNA)

RNA transcripts (many possible variants)

RNA-seq reads

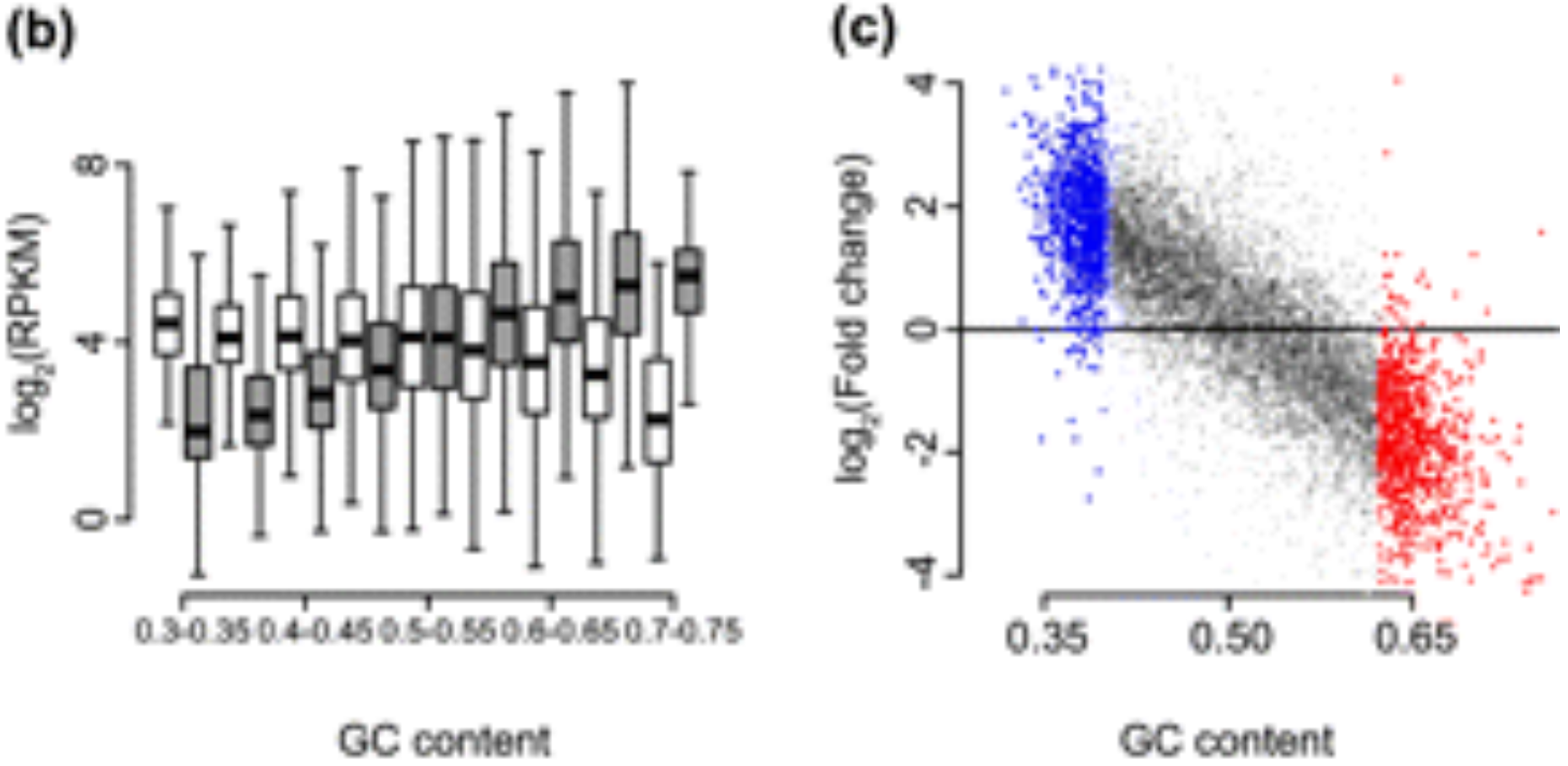


Genome (DNA)

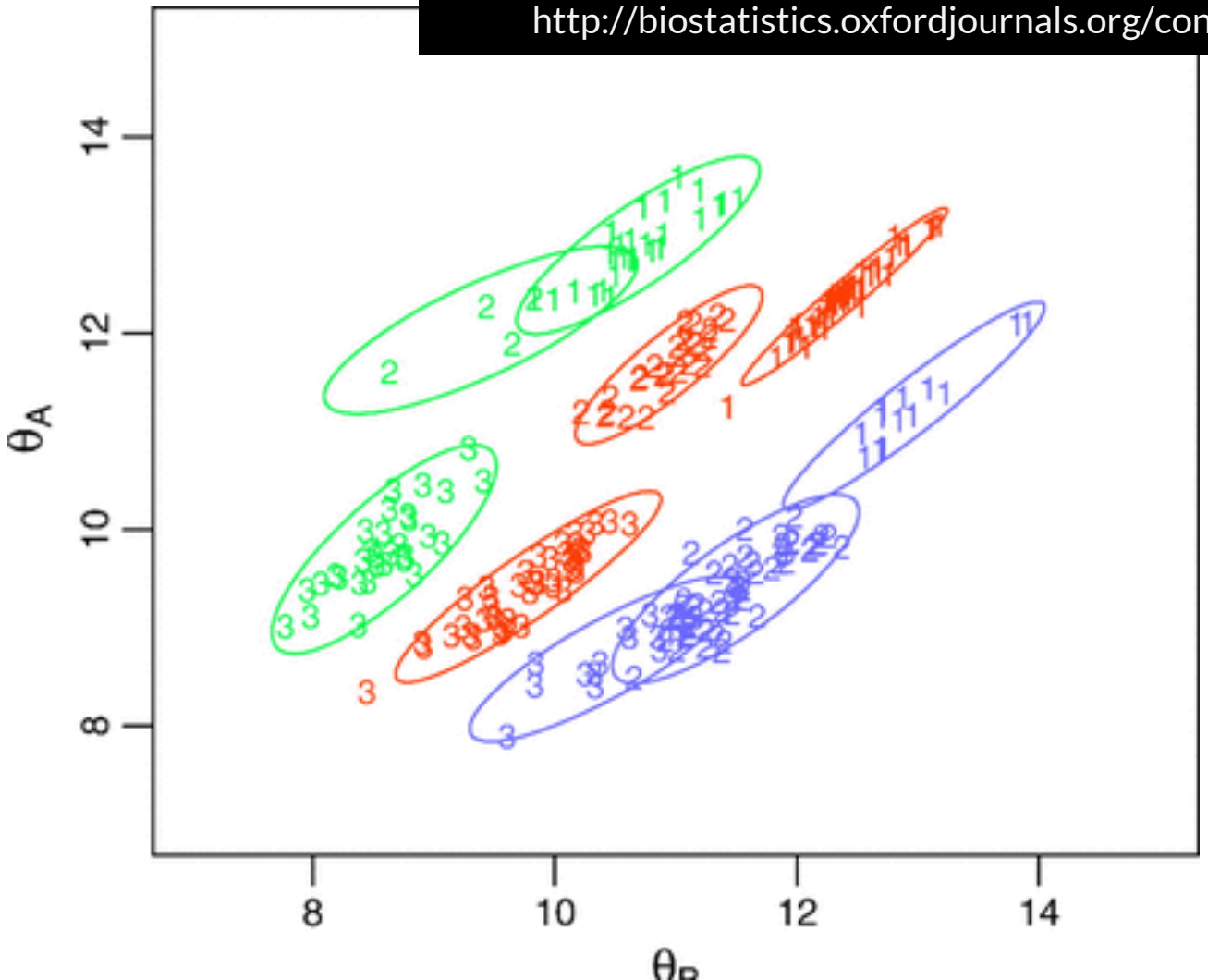


expression = 24

http://biostatistics.oxfordjournals.org/content/early/2012/01/24/biostatistics.kxr054.long



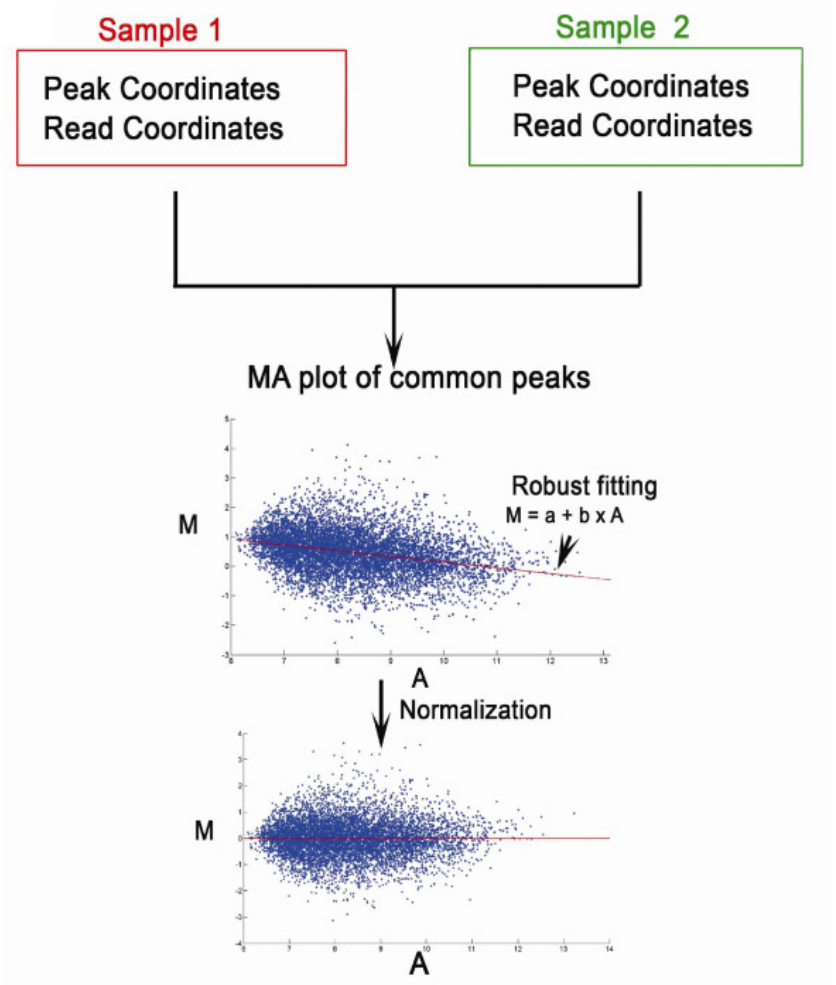
http://biostatistics.oxfordjournals.org/content/8/2/485.long



Normalization

- Remove technological biases
- Make samples comparable

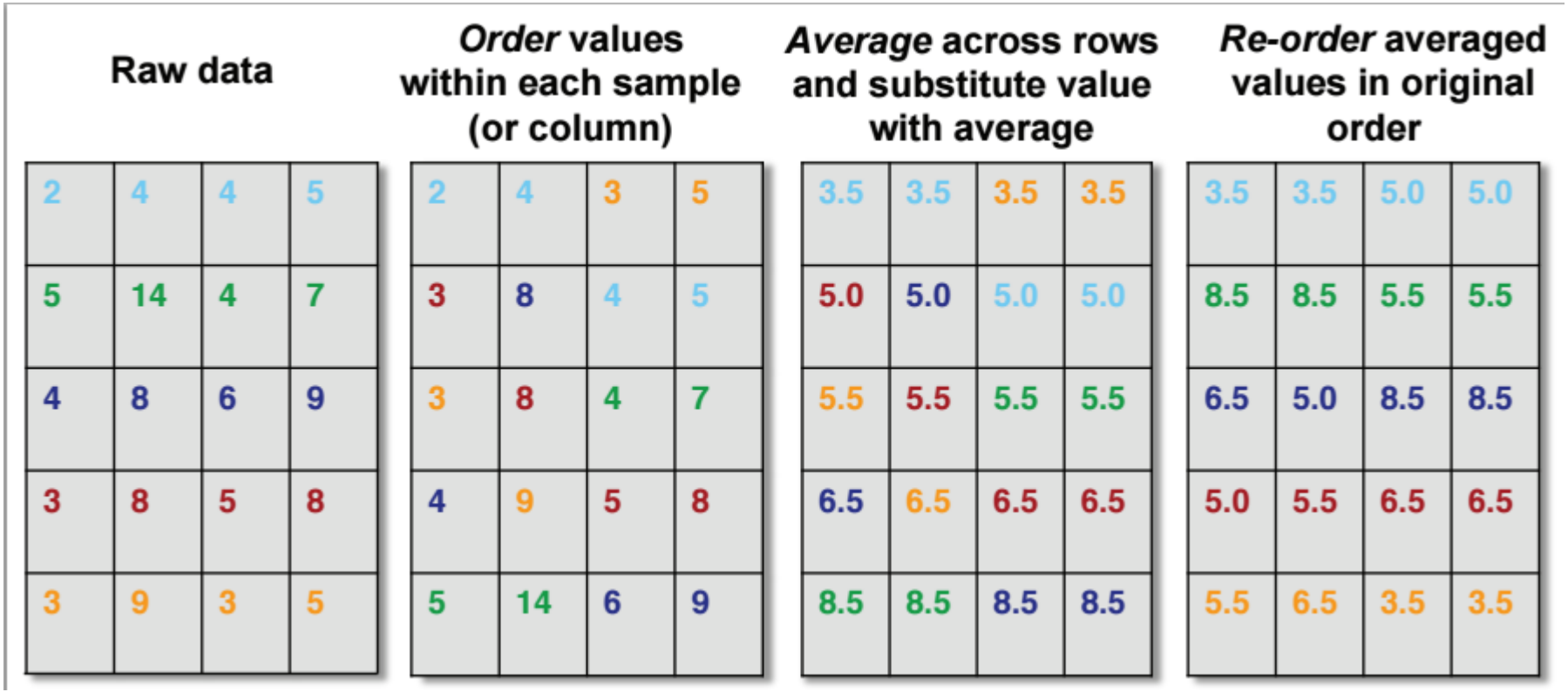
http://www.genomebiology.com/2012/13/3/R16/abstract



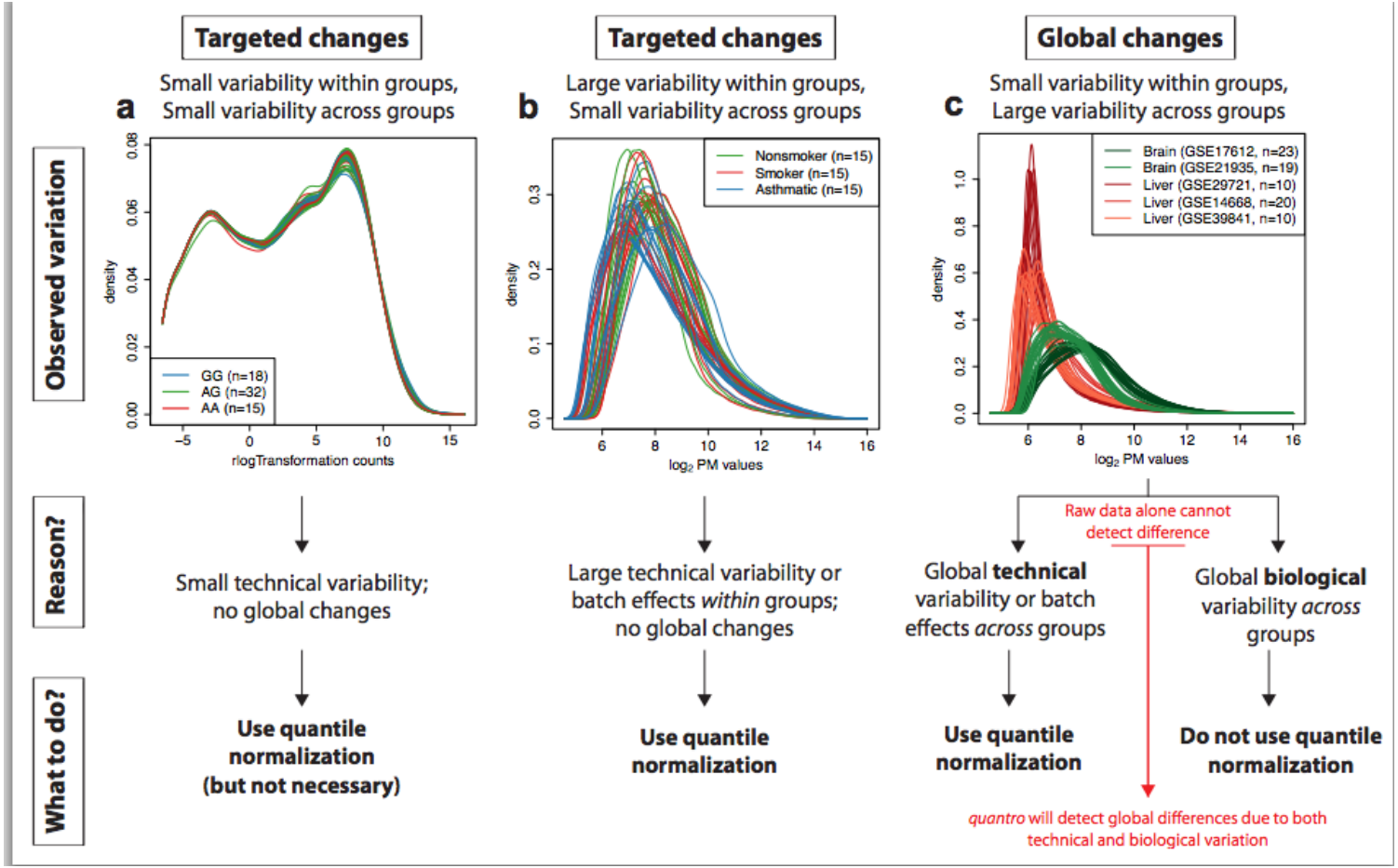
Quantile normalization

- Most common technique
- Bulk distributions exactly the same

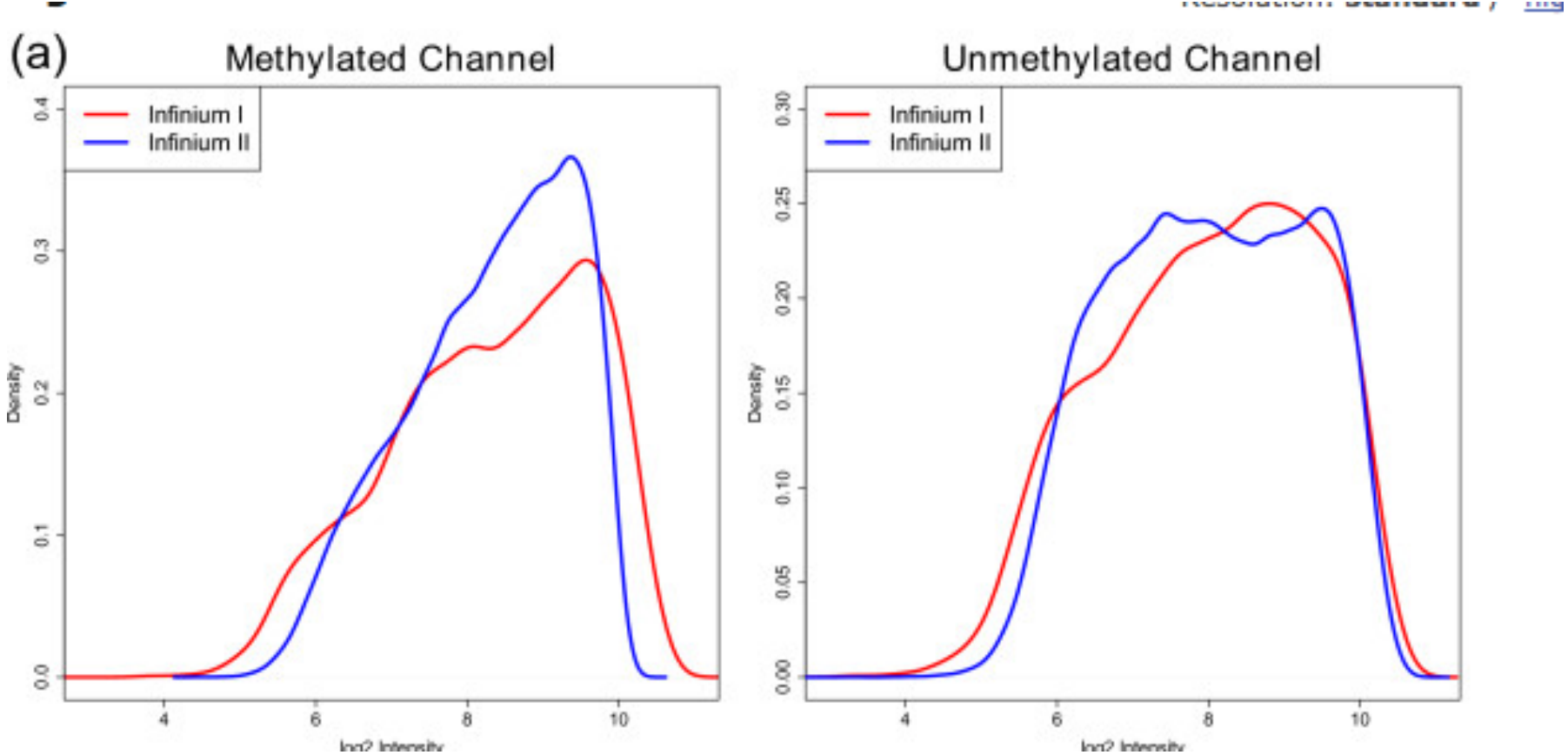
http://www.genomebiology.com/2015/16/1/117



http://www.genomebiology.com/2015/16/1/117



http://www.genomebiology.com/2012/13/6/r44



Notes and further reading

- Preprocessing and normalization are highly platform/problem dependent
- In general check to make sure there aren't bulk differences between samples, especially due to technology
- Bioconductor workflows are a good place to start: https://www.bioconductor.org/help/workflows/

http://www.nature.com/nature/journal/v487/n7408/full/487427a.html

“First, researchers starting out in genomics must keep in mind that interesting outliers — that is, results that deviate significantly from the sample — will inevitably contain a plethora of experimental or analytical artefacts.”