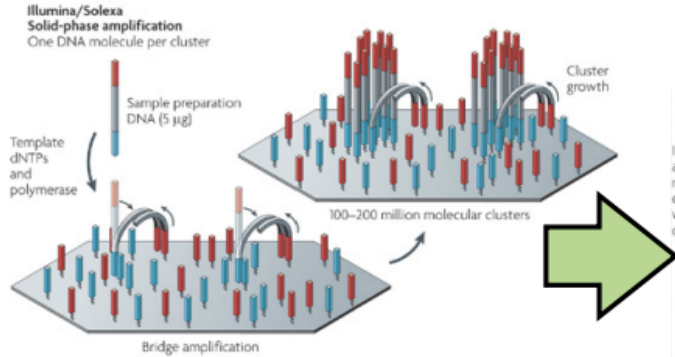


Pre-processing and normalization

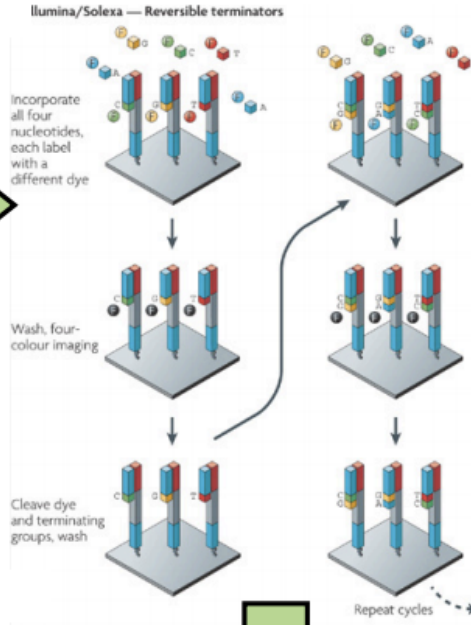
Jeff Leek

@jtleek

www.jtleek.com



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



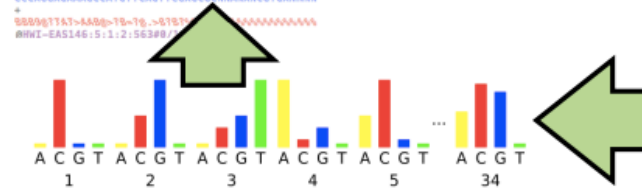
```

@HWI-EAS146:5:1:1:961#0/1
TCCGAGGCCAACCGAGGCTCCGCGCGCTGNNNNNNNNNNNNNNNN
+
BBBB-A7BB:0BBBBA-BA-A*****
@HWI-EAS146:5:1:1:1595#0/1
TCAGGAAGCAGGAAGAGCTGGTCAGCAGGNNNNNNNNNNNNNNNN
+
09B0B-; BAA-0AB9-1-1*****
@HWI-EAS146:5:1:1:1848#0/1
CTGGACTGCATCCTACCACCAACTCTCCAAANNNNNNNNNNNNNN
+
A-B7A7->BB-A>79-;<; :>747*****
@HWI-EAS146:5:1:1:1719#0/1
CAGCATCTGGTATTATTGTAACCTCCGCTCANNNGHTNAAGNNNN
+
BCC7+-B=7BB5=ABA7B6BBB4BB7B*****
@HWI-EAS146:5:1:2:947#0/1
CCCAGGAAAGCCATGTTCACTTCAGCGCANNANACTGANNNNN
+
BBD9@TAT7-AA0@-19-10->6T0*
@HWI-EAS146:5:1:2:563#0/1

```

name
sequence
quality scores

x 100s of millions



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

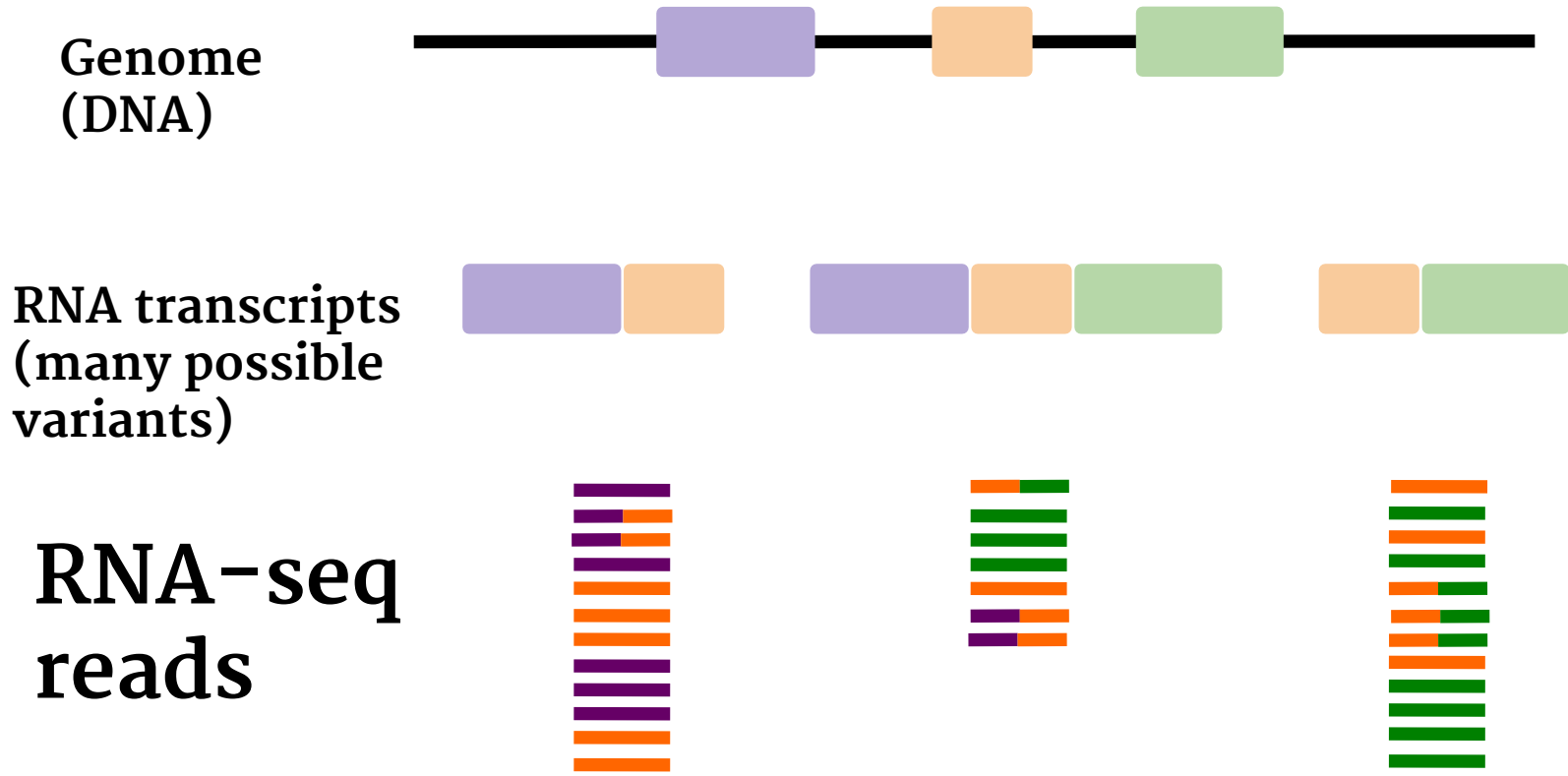


Source: Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010

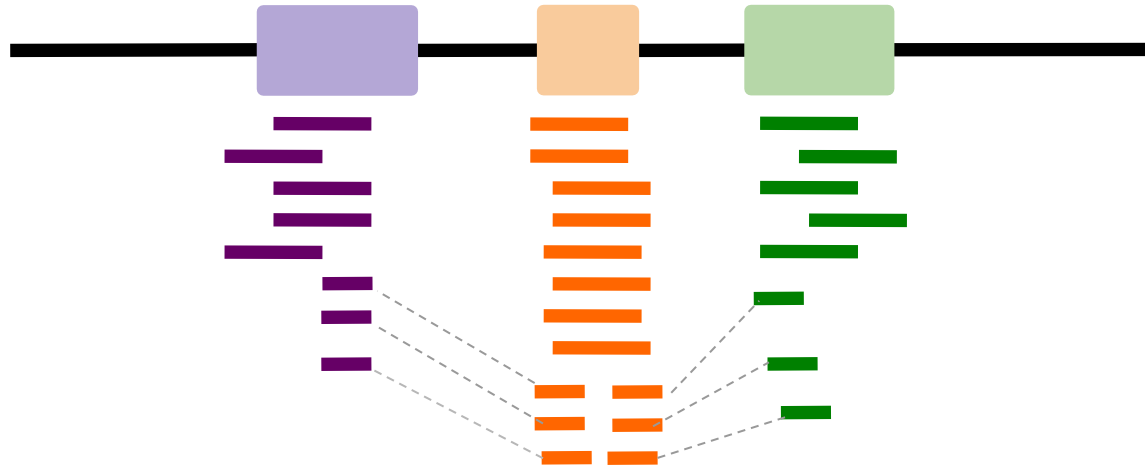
Preprocessing

Convert raw data to “processed”

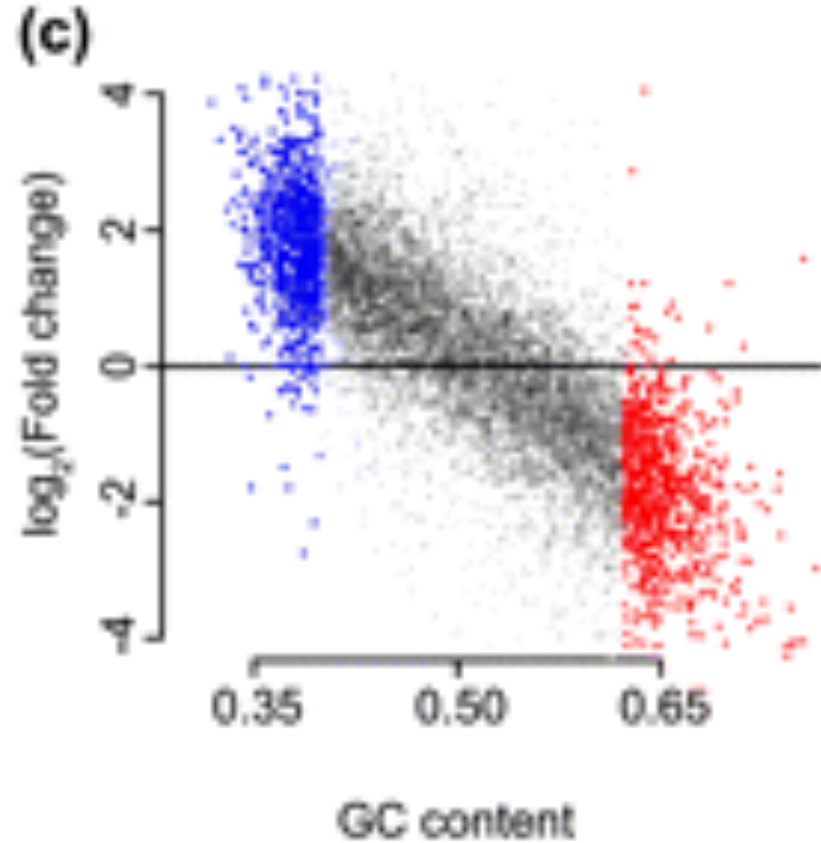
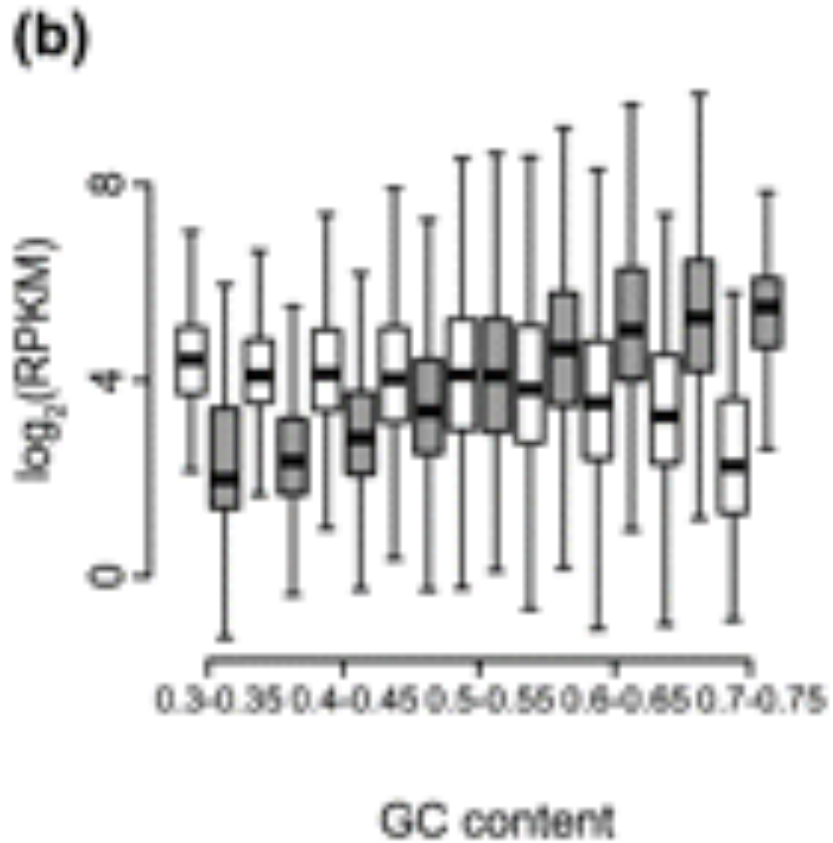
Try to remove technological artifacts

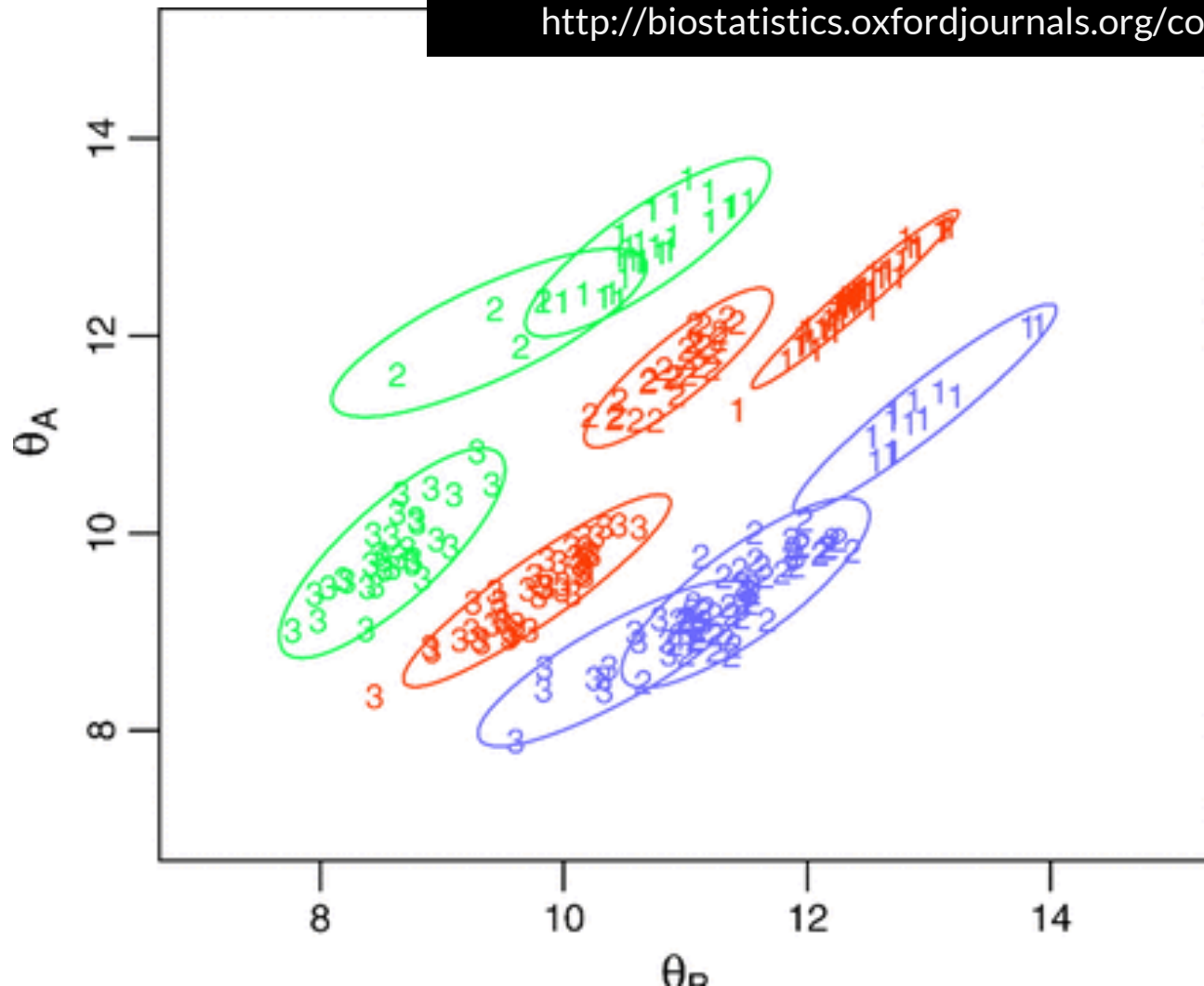


Genome
(DNA)



expression = 24





Normalization

Remove technological biases

Make samples comparable

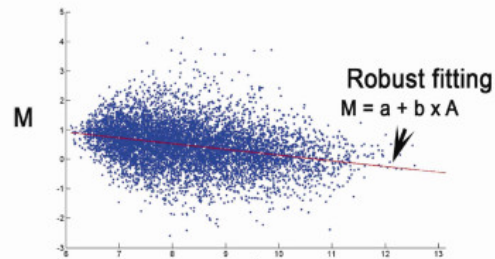
Sample 1

Peak Coordinates
Read Coordinates

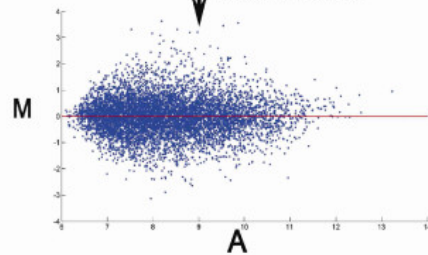
Sample 2

Peak Coordinates
Read Coordinates

MA plot of common peaks



Normalization



Quantile normalization

Most common technique

Bulk distributions exactly the same

Raw data

2	4	4	5
5	14	4	7
4	8	6	9
3	8	5	8
3	9	3	5

Order values within each sample (or column)

2	4	3	5
3	8	4	5
3	8	4	7
4	9	5	8
5	14	6	9

Average across rows and substitute value with average

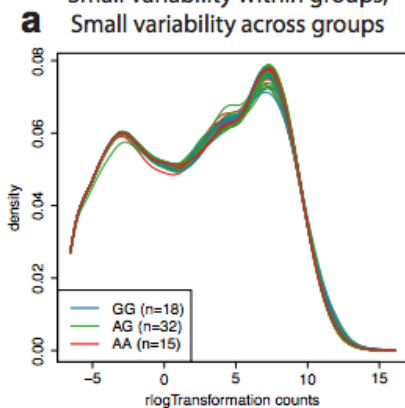
3.5	3.5	3.5	3.5
5.0	5.0	5.0	5.0
5.5	5.5	5.5	5.5
6.5	6.5	6.5	6.5
8.5	8.5	8.5	8.5

Re-order averaged values in original order

3.5	3.5	5.0	5.0
8.5	8.5	5.5	5.5
6.5	5.0	8.5	8.5
5.0	5.5	6.5	6.5
5.5	6.5	3.5	3.5

Targeted changes

Small variability within groups,
Small variability across groups



Observed variation

Reason?

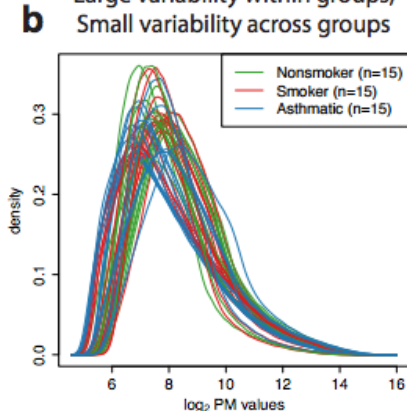
What to do?

Small technical variability;
no global changes

Use quantile
normalization
(but not necessary)

Targeted changes

Large variability within groups,
Small variability across groups

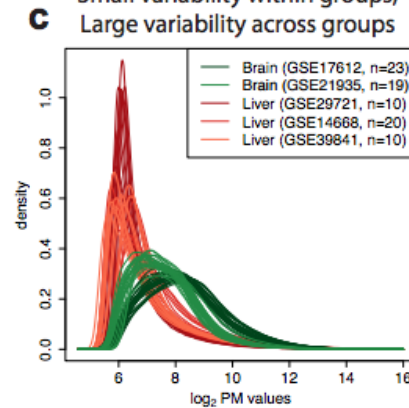


Large technical variability or
batch effects *within* groups;
no global changes

Use quantile
normalization

Global changes

Small variability within groups,
Large variability across groups



Global **technical**
variability or batch
effects *across* groups

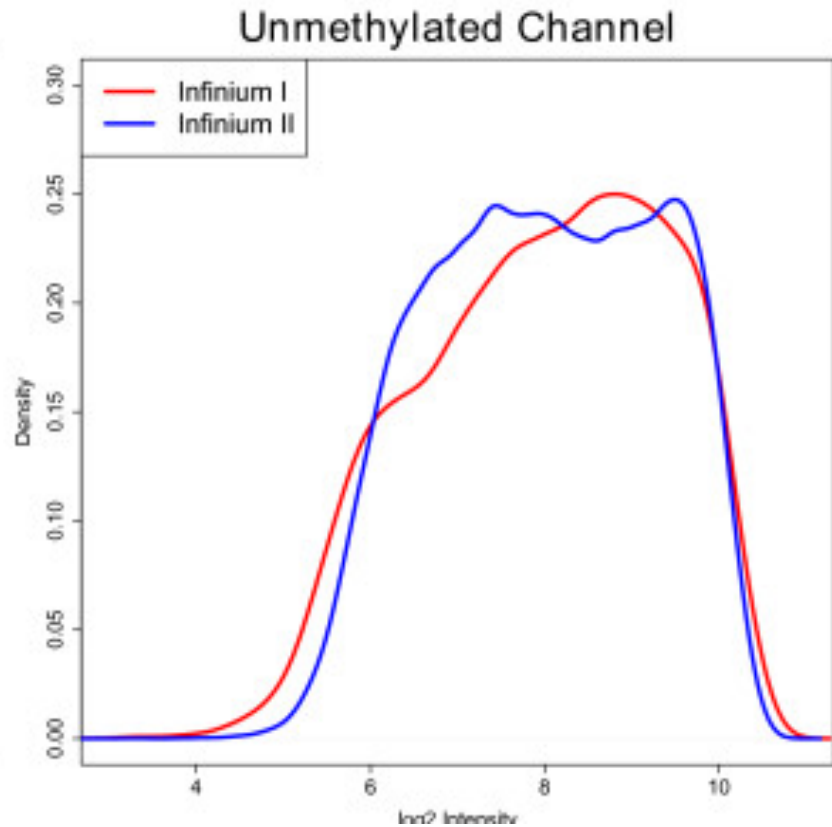
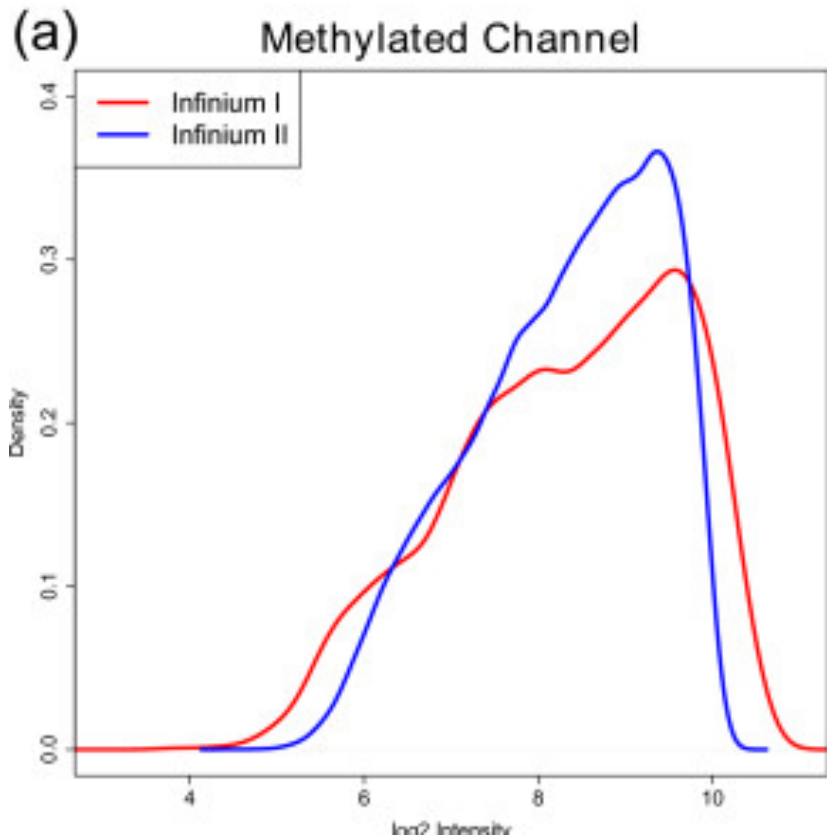
Use quantile
normalization

Global **biological**
variability *across*
groups

Do not use quantile
normalization

Raw data alone cannot
detect difference

quantro will detect global differences due to both
technical and biological variation



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/>

“ 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. ”