Assessing Sequence and Microarray Data Quality

Prat Thiru

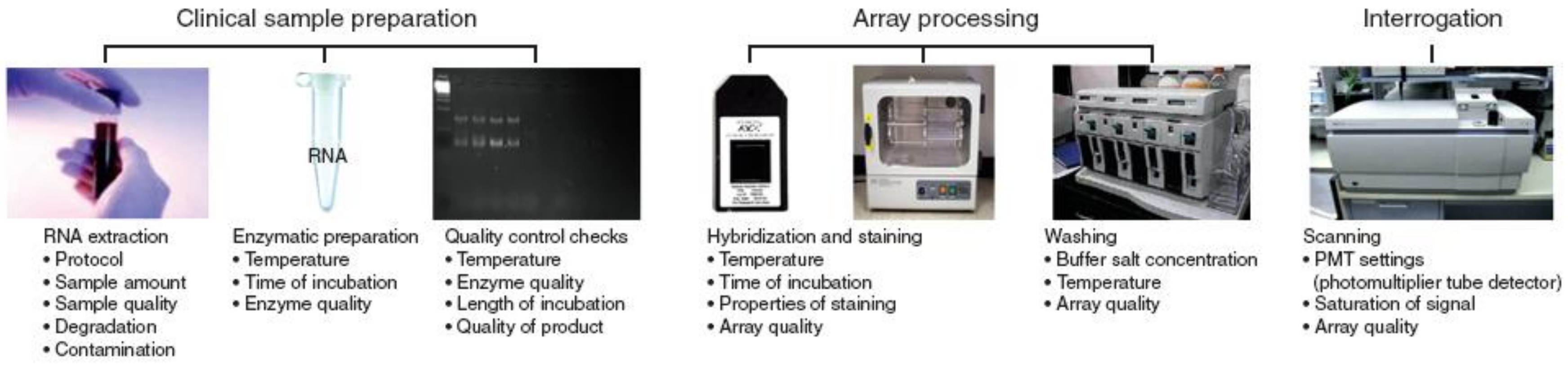
Outline

- Introduction
- Examples and Interpreting QC Reports
- Batch Effects
- Tools available for QC
  - Microarray
  - Short-Reads
- Work Flow

Consequences of not Assessing the Data

- Increased variability and decreased power to detect biological significance
- Waste of resources: cost and time
- Study is not reproducible
- Downstream analysis can be incorrect
  - ➢Microarrays: Normalization fails to remove noise
  - ➢Short-Reads: reads fail to map or align

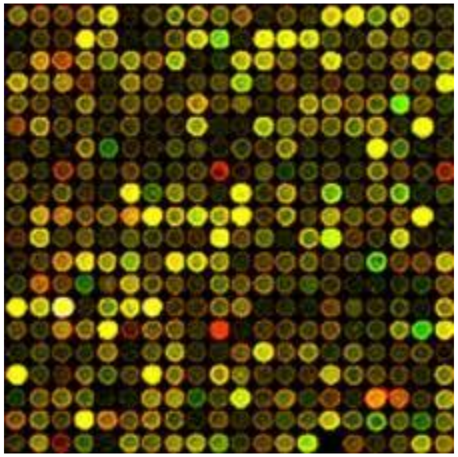
Data Integrity Needed at Multiple Steps

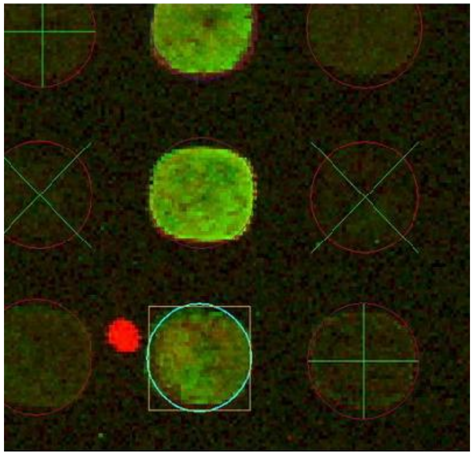


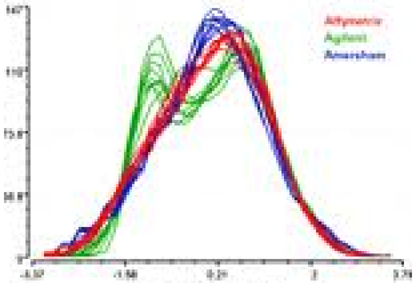
Ji, H. and Davis, R.W. *Data quality in genomics and microarrays.* Nature Biotechnology 24:9 (2006)

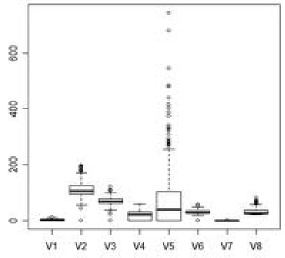
Array Data

- Measure intensity or pixel values
- Plot or analyze the intensity values to assess data quality
- Distribution of intensities should be similar since most genes are not differentially expressed



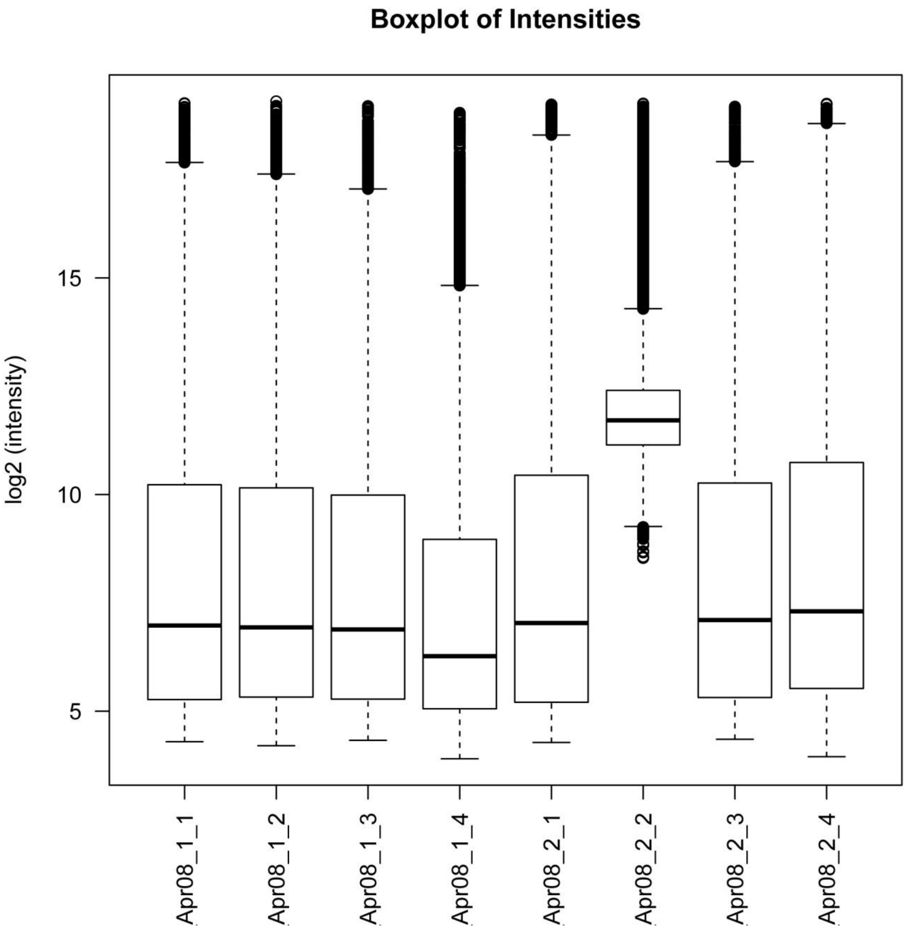






Microarray: Box Plots
Agilent One-Color

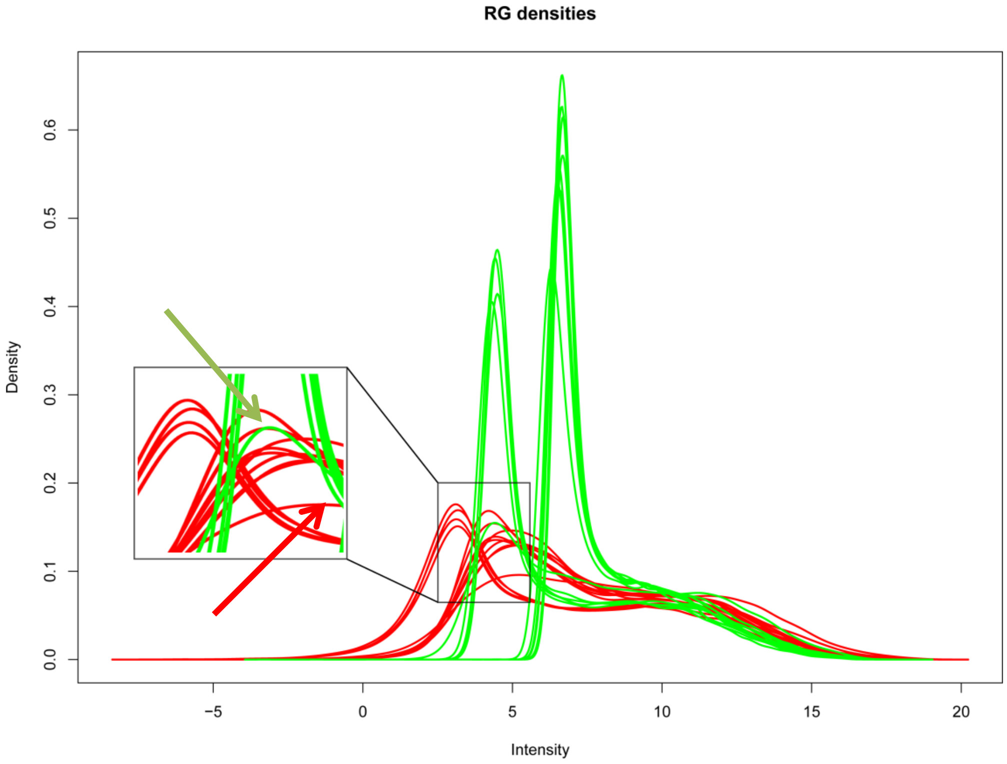
- Box plots of intensity values shows distribution across arrays
- Array Apr08_2_2 (on figure) has a dramatically different distribution compared to other arrays



Boxplots can be created using R boxplot *command or using the Bioconductor package arrayQualityMetrics*

Microarray: Density Plot
Agilent Two-Color

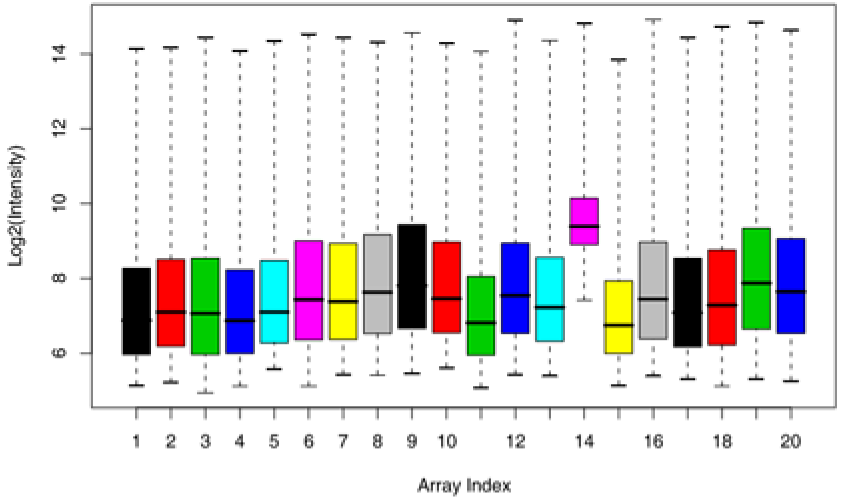
- Density plot, a smoothed-histogram, shows intensity distribution of each array.
- Data from two experiments can be seen by the two distinct (red and green) peaks (on figure). A single (red and green shown by arrows) peak shows a problematic array (inset).

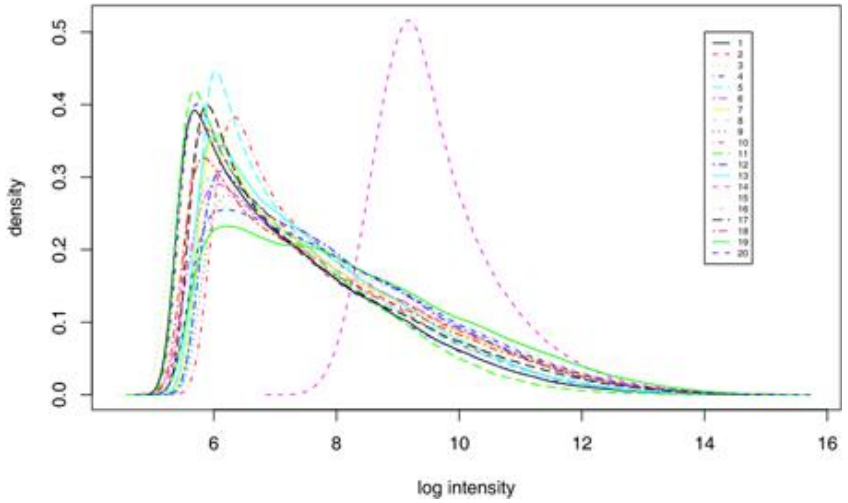


Density plot can be created using R plotDensities *command from limma package or using the Bioconductor package arrayQualityMetrics*

Microarray: Box Plot and Density Plot

- Combining both box plot and density plot shows arrays that need to be carefully examined, and if they should be included in further analysis





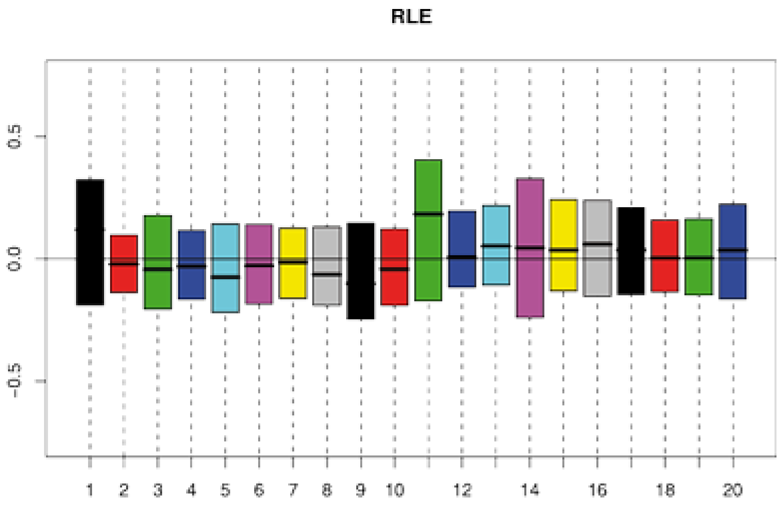
https://www.genevestigator.com/userdocs/manual/qc.html

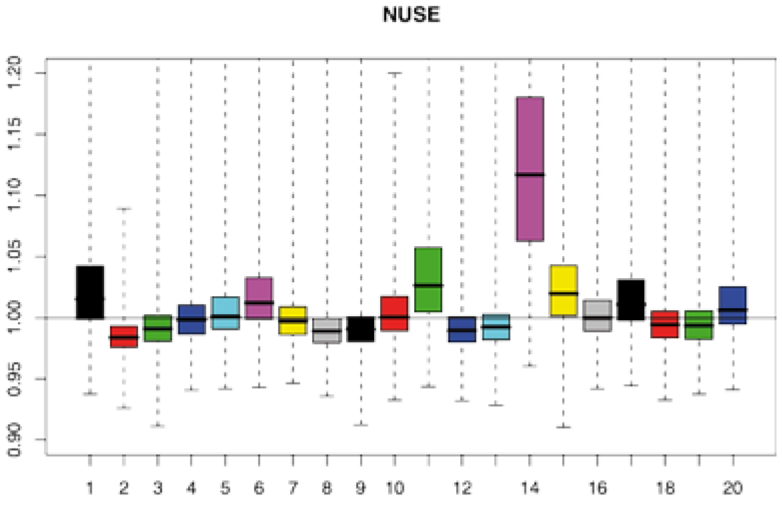
Microarray: RLE and NUSE

- Relative Log Expression (RLE): Comparison of probeset expression value on each array across the median expression value for that probeset on all arrays.
- Normalized Unscaled Standard Error (NUSE): normalized standard error estimates from the probe level model (PLM)
- Available for Affymetrix, using the commands NUSE and RLE from the package affyPLM

Microarray: RLE and NUSE

- RLE: Expression of most genes remain the same, RLE values should be close to 0
- NUSE: if a median standard error of 1 is used, then check if NUSE values are close to 1





https://www.genevestigator.com/userdocs/manual/qc.html

Sequence Data

- Reads have sequencing quality information
- Fastq format:

```
@ILLUMINA-F6C19_0048_FC:5:1:12440:1460#0/1   ←   Header (if Paired-End (PE) ➔ /1)
GTAGAACTGGTACGGACAAGGGGAATCTGACTGTAG         ←   Sequence
+ILLUMINA-F6C19_0048_FC:5:1:12440:1460#0/1   ←   Description
hhhhhhhhhhghhhhhhhehhhedhhhhfhhhhhh          ←   Quality Values
```

- Examine the quality values to assess sequence data

Sequence Quality

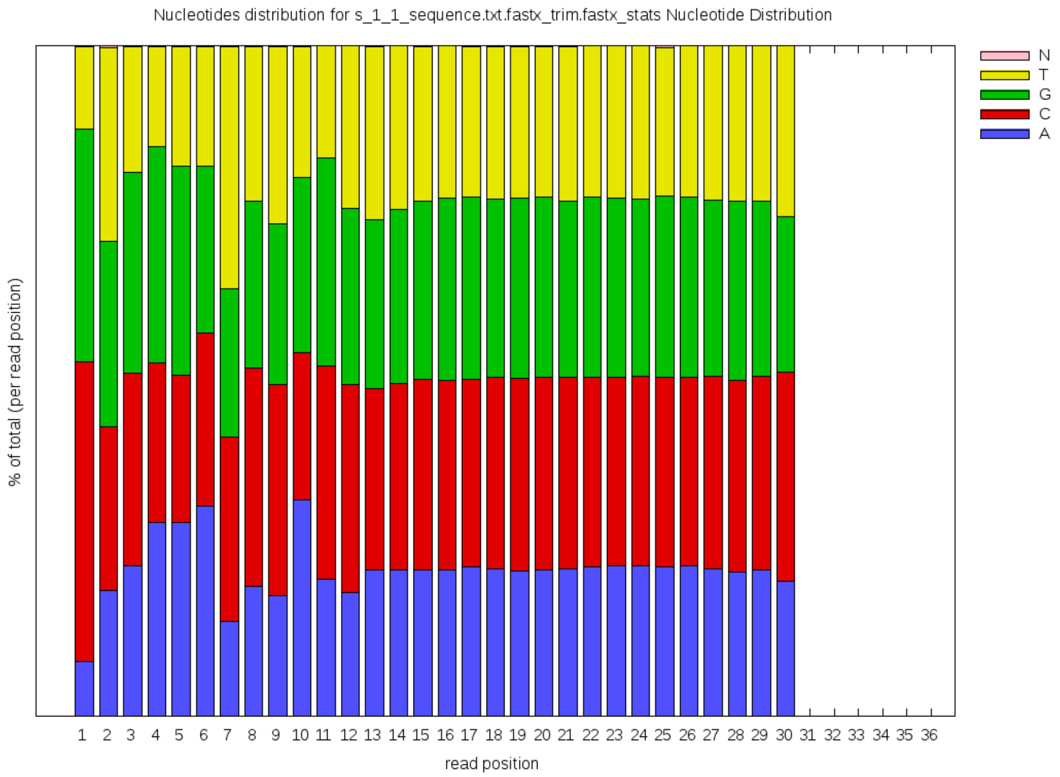
- Quality values in fastq files are ASCII-encoded using 64 to 126 (Illumina Pipeline 1.3)

eg. Solexa quality score h = 104 – 64 = 40

where 104 is the ASCII value for “h” and 64 is the offset

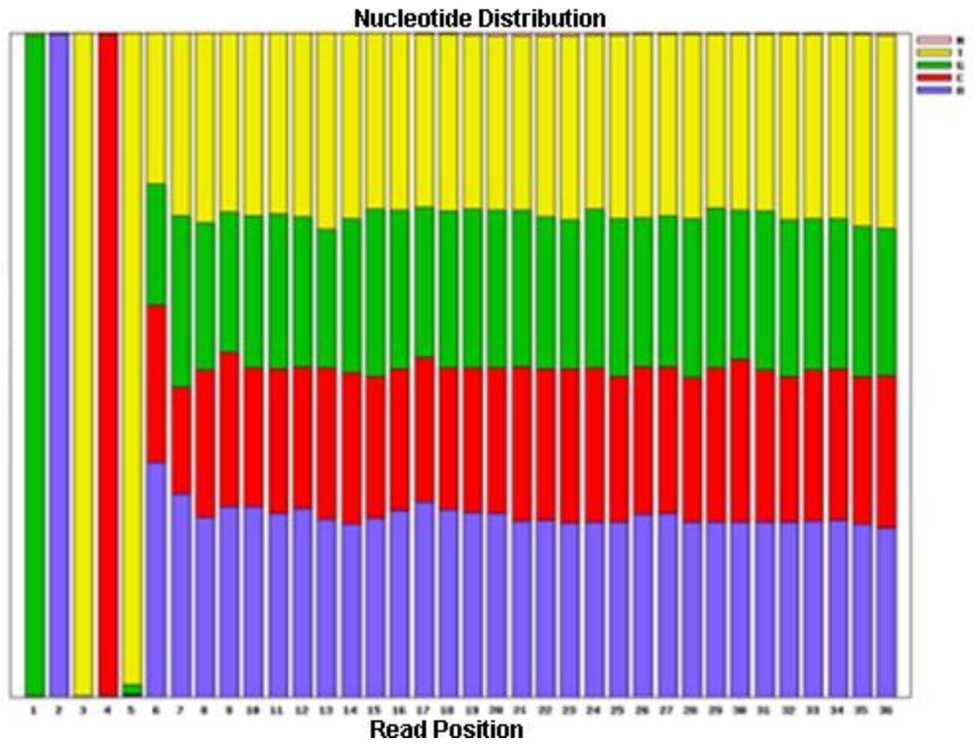
Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.90%
40	1 in 10000	99.99%

Sequence: Nucleotide Distribution and Barcode



Fastx Toolkit fastx_nucleotide_distribution_graph.sh

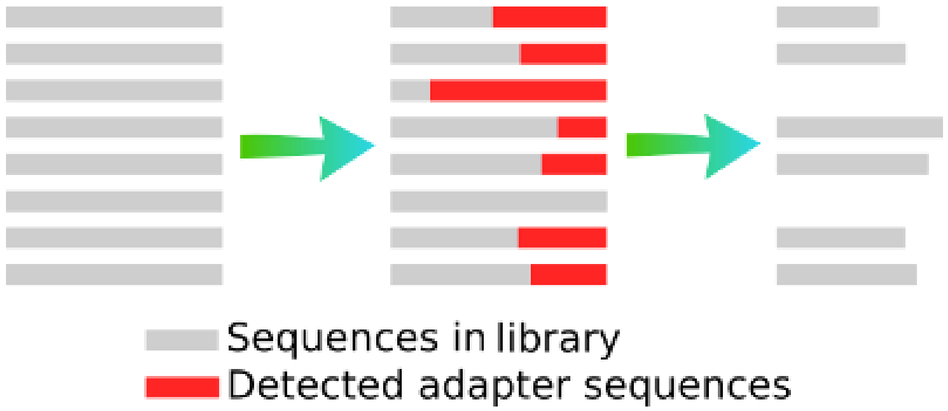
The following chart clearly shows the barcode used at the 5'-end of the library: **GATCT**

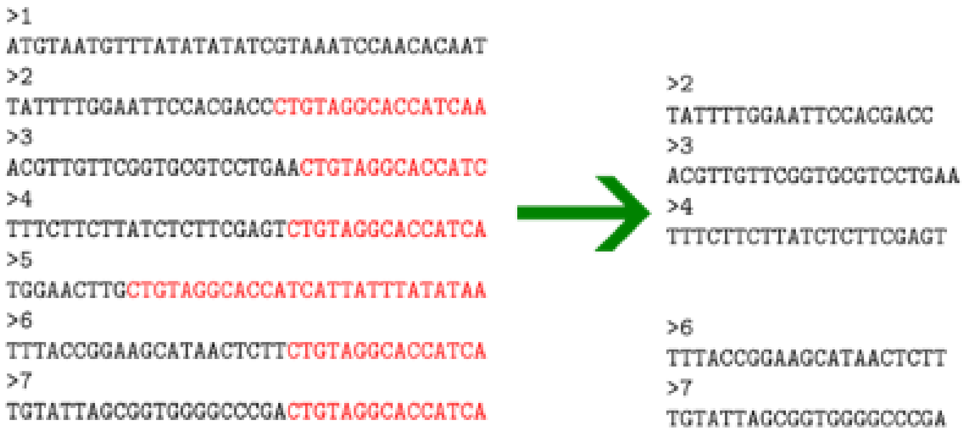


http://main.g2.bx.psu.edu/static/fastx_icons/fastq_nucleotides_distribution_1.png

Sequence: Adapter or Linker

- Clip adapters from 3'-end and ensure reads are at least a certain minimum length
- Sequence 1 was discarded since it wasn't clipped (ie. no adapter sequence) and Sequence 5 was discarded since its length after clipping was too short (see figure)





```
>1
ATGTAATGTTTATATATATCGTAAATCCAACACAAT
>2
TATTTTGGAATTCCACGACCCTGTAGGCACCATCAA
>3
ACGTTGTTCGGTGCGTCCTGAACTGTAGGCACCATC
>4
TTTCTTCTTATCTCTTCGAGTCTGTAGGCACCATCA
>5
TGGAACTTGCTGTAGGCACCATCATTATTTATATAA
>6
TTTACCGGAAGCATAACTCTTCTGTAGGCACCATCA
>7
TGTATTAGCGGTGGGGCCCGACTGTAGGCACCATCA
```

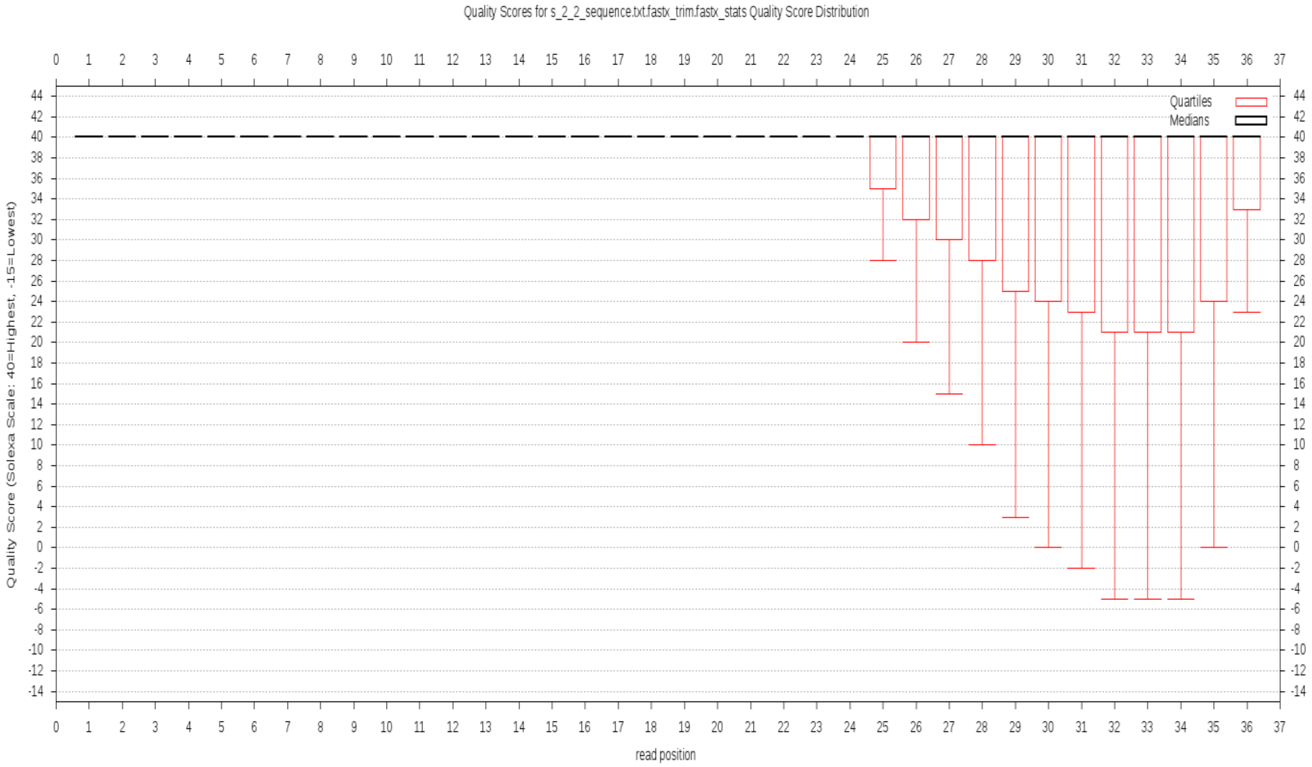
```
>2
TATTTTGGAATTCCACGACC
>3
ACGTTGTTCGGTGCGTCCTGAA
>4
TTTCTTCTTATCTCTTCGAGT

>6
TTTACCGGAAGCATAACTCTT
>7
TGTATTAGCGGTGGGGCCCGA
```

Fastx Toolkit fastx_clipper

http://main.g2.bx.psu.edu/static/fastx_icons/fastx_clipper_illustration.png
http://main.g2.bx.psu.edu/static/fastx_icons/fastx_clipper_example.png

Sequence: Quality Score Distribution

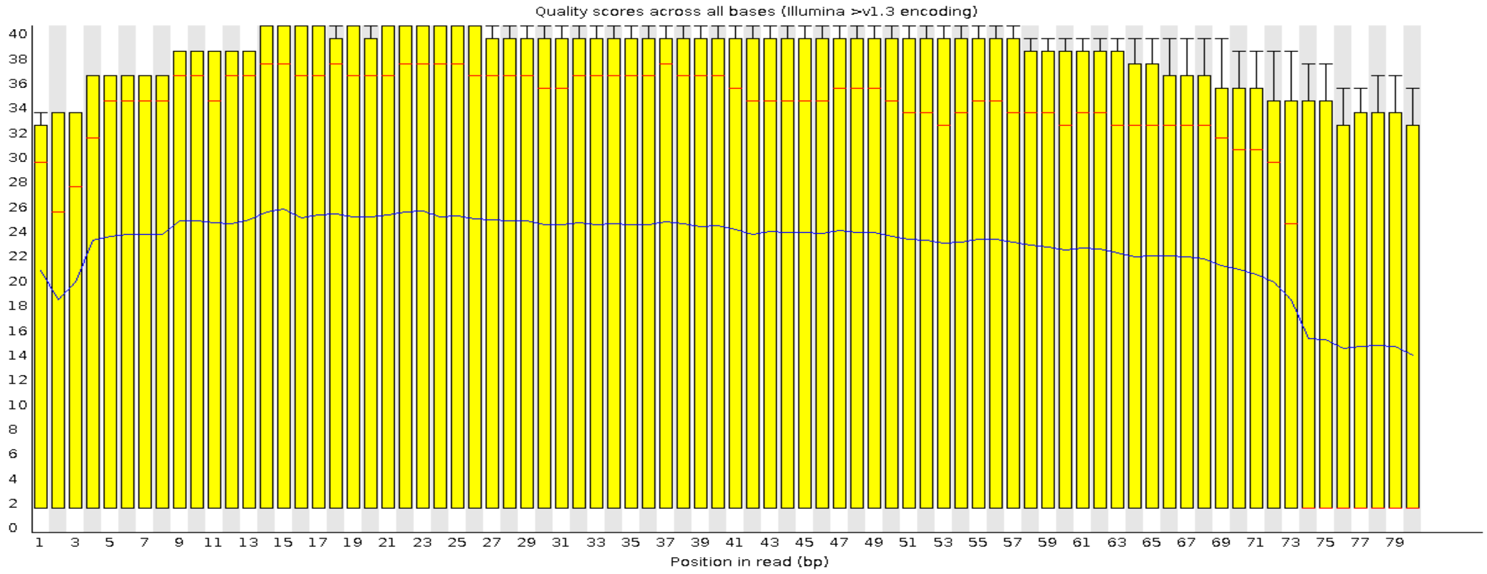
- The error rates towards the end of the reads increase due to sequencing cycles
- Low quality towards other parts of the reads indicates reads that might need to be trimmed, or completely removed



Fastx Toolkit fastq_quality_boxplot_graph.sh

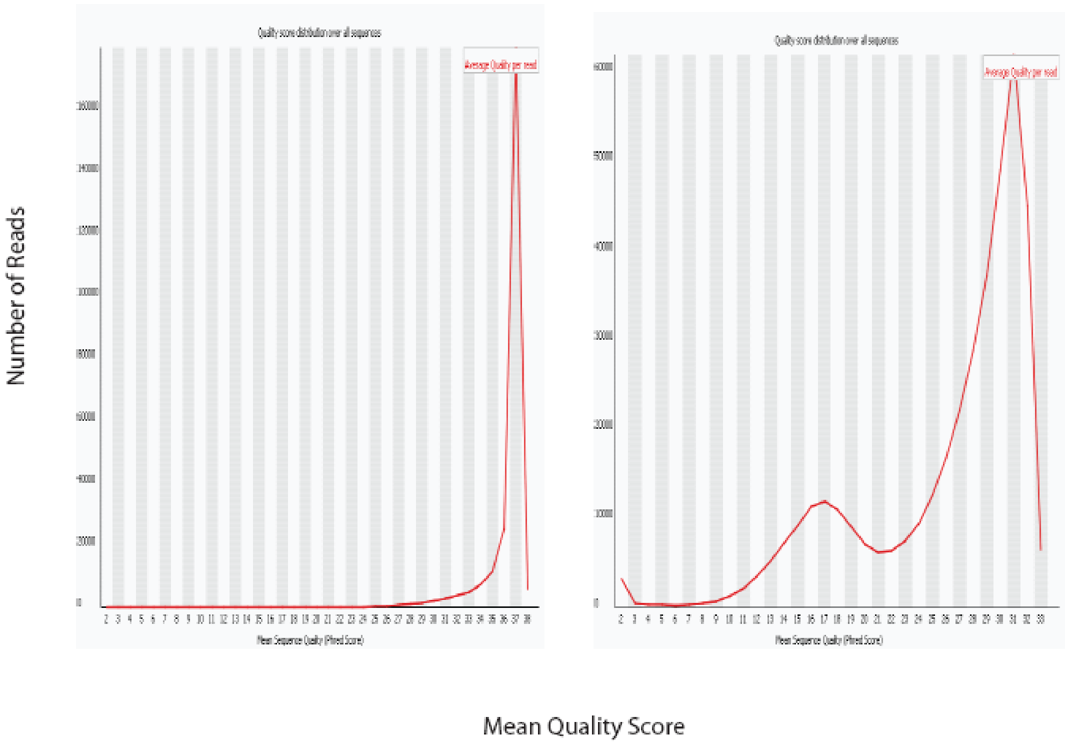
Sequence: Quality Score Distribution

- An example where all positions in the read had questionable quality



Sequence: Average Quality per Read Distribution

- Distribution should have a single peak towards high quality
- A bi-modal graph (on the right figure) shows problematic reads



FastQC Tool

http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/

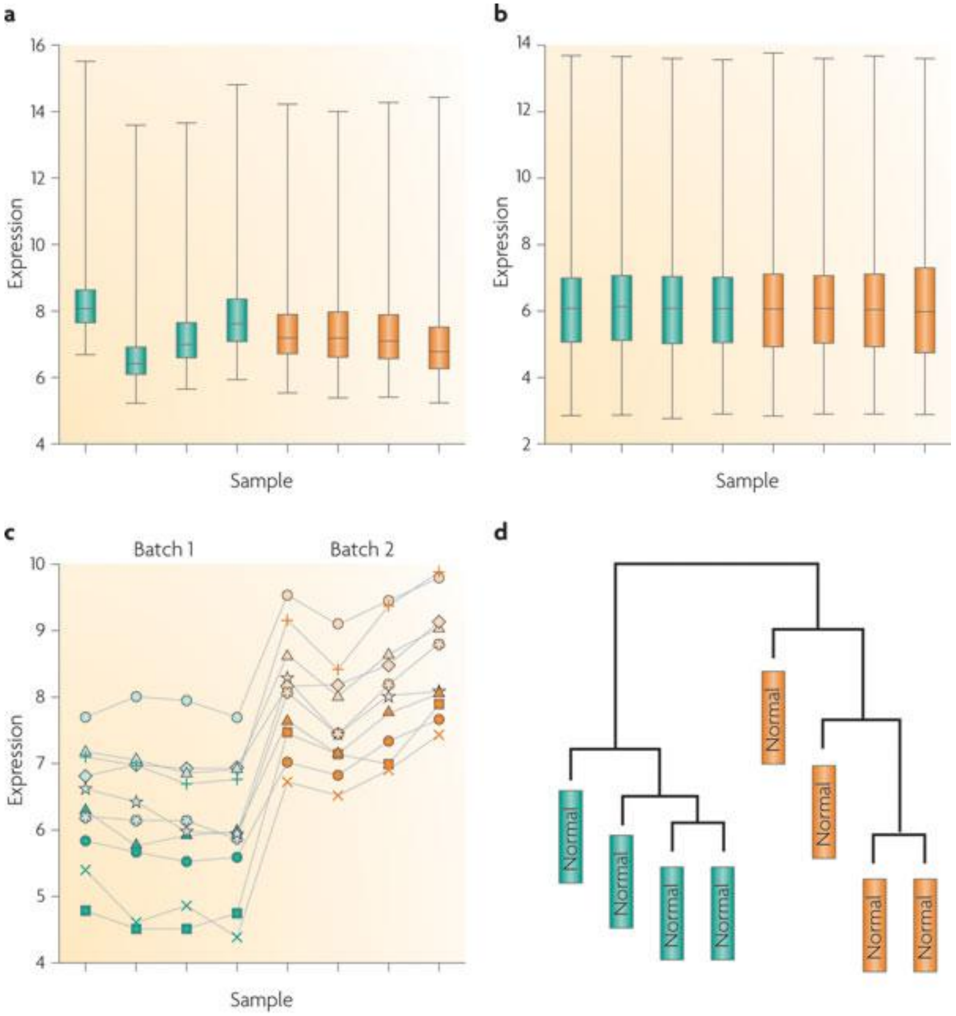
Batch Effects

- Sub-groups of measurements that have qualitatively different behavior across conditions and are unrelated to the biological or scientific variables in a study
- Batch effects may occur if a subset of experiments was run on different days, technicians, labs, etc.
- Normalization does not completely remove batch effects

Irizarry, A.R, et al. *Tackling the widespread and critical impact of batch effects in high-throughput data* Nature Reviews Genetics 11:733 (2010)

Batch Effects Example

- Studies can be confounded by batch effects: an extraneous variable correlates with both the outcome and an independent variable of interest (eg. gene expression)
- Not easy to find!

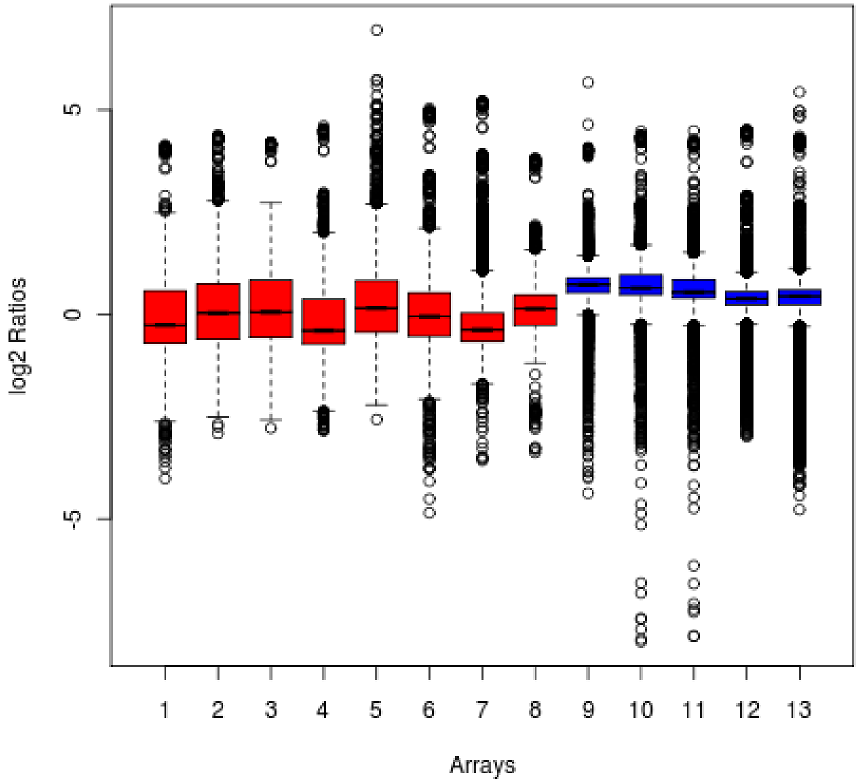


Nature Reviews | Genetics

Irizarry, A.R, et al. *Tackling the widespread and critical impact of batch effects in high-throughput data* Nature Reviews Genetics 11:733 (2010)

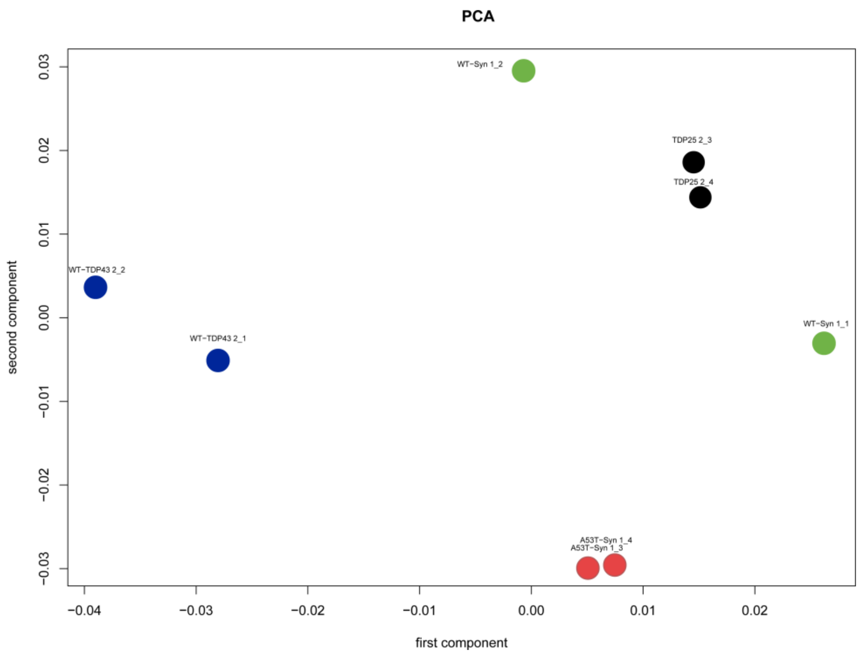
Agilent Two-Color Box Plots

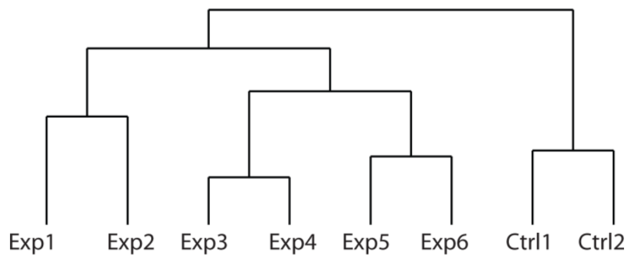
- Box plots of raw data by combining two experiments shows different distributions
- Downstream analysis should consider if differences seen are of biological significance or simply due to two different experiments.
- Normalization may not completely remove differences due to different experiments



Finding Batch Effects

- Meta data: good documentation of date, time, who carried out the experiment, etc.
- Good experimental design
- Principal Component Analysis (PCA) and unbiased clustering may show batch effects





PCA (top figure)
Dendogram from Clustering (bottom figure)

Assessing Data:
Downstream Analysis

- Microarray
  - MA Plot
  - Volcano Plot
  - Check if normalization worked
- Sequence
  - Evaluate alignment: basic stats (eg. percent mapped, insert size distribution for PE data, etc.)
  - Examine unmapped reads
- Clustering
- Principal Component Analysis (PCA)

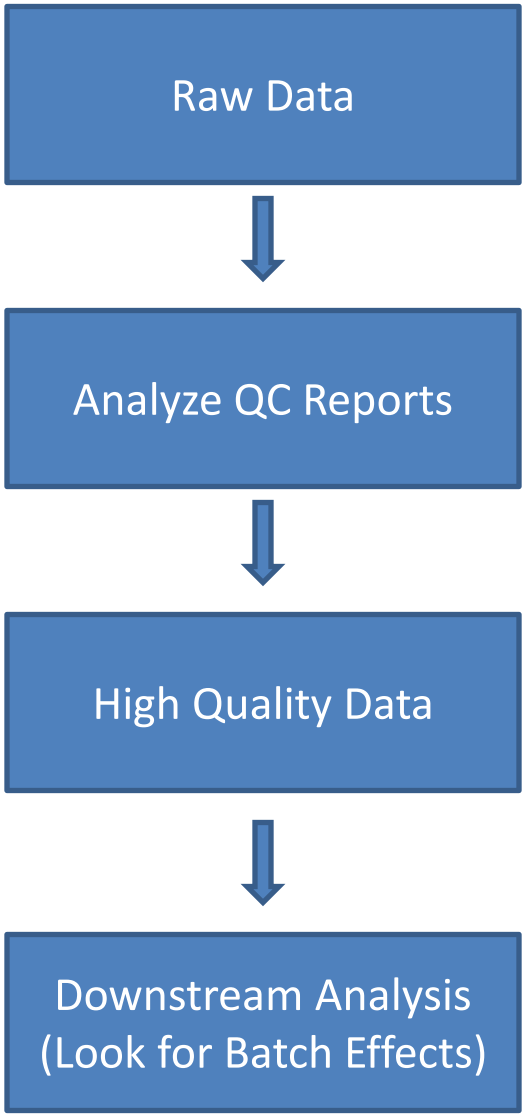
http://iona.wi.mit.edu/bio/education/hot_topics/diff_exp/Hot_Topics_-_Differential_Expression_2008_color.pdf
http://jura.wi.mit.edu/bio/education/bioinfo2007/arrays/

QC Tools

Data	Tool	GUI	Command Line	Website
Array	affyPLM		*	Bioconductor: http://www.bioconductor.org
Array	arrayQualityMetrics		*	Bioconductor: http://www.bioconductor.org
NGS	Fastx Toolkit	*	*	http://hannonlab.cshl.edu/fastx_toolkit Or Galaxy: http://main.g2.bx.psu.edu/
NGS	FastQC	*	*	http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc Or Galaxy: http://main.g2.bx.psu.edu
NGS	ShortRead		*	Bioconductor: http://www.bioconductor.org

Next-Gen Sequencing (NGS)
Graphical User Interface (GUI)

Work Flow


Raw Data

Analyze QC Reports

High Quality Data

Downstream Analysis (Look for Batch Effects)

- eg. Affy .CEL files, Solexa .fastq files
- Use available QC tools for array or sequence data
- Filter data: remove low quality data (eg. reads, samples, etc.)
- Normalization, clustering, mapping/alignment, etc.

More Information

- Whitehead Genome Technology Core:

http://jura.wi.mit.edu/genomecorewiki/index.php/SequencingQC

http://jura.wi.mit.edu/genomecorewiki/index.php/MicroarrayQC

- Microarray Quality Control (MAQC)

http://www.nature.com/nbt/focus/maqc/index.html

- Standards, Guidelines and Best Practices for RNA-Seq (from ENCODE)

http://encodeproject.org/ENCODE/protocols/dataStandards/ENCODE_RNAseq_Standards_V1.0.pdf

- BaRC Standard Operating Procedures (SOP)

https://gir.wi.mit.edu/trac/wiki/barc/SOPs