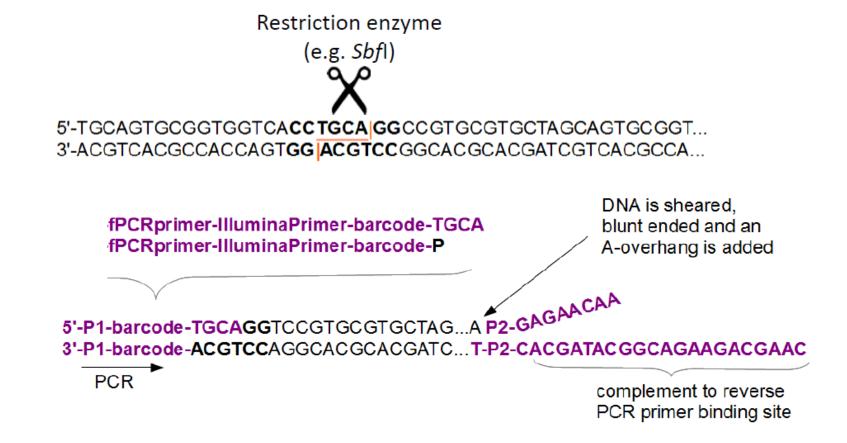
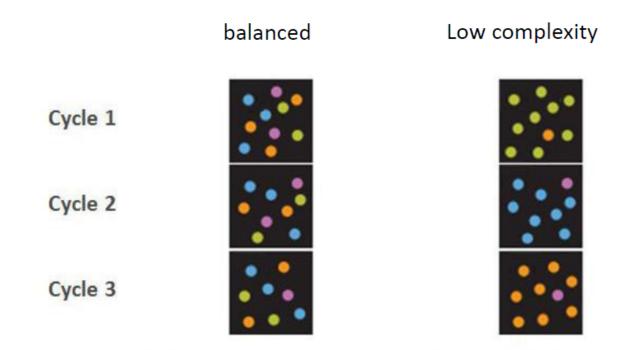
## RAD/GBS



Each read: will start with the barcode, then the restriction site, then a variable sequence

### Issues with cluster identification

Due to low complexity at the beginning of the sequence, Illumina cannot distinguish if a signal comes from one or two clusters



## How to minimize the problem

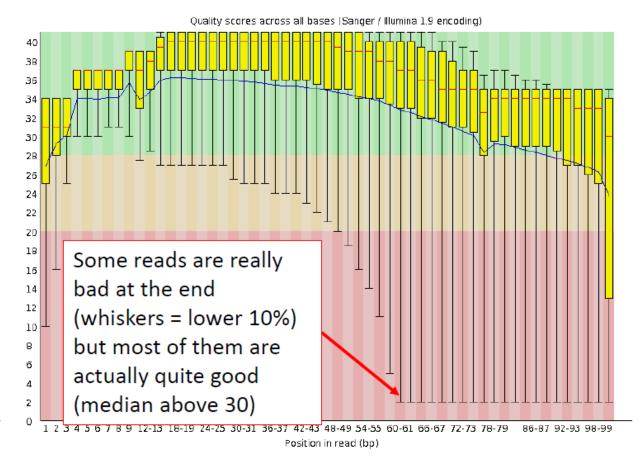
- Use barcodes of different lengths to shift the restriction enzyme cut site
- Add PhiX virus DNA to the RAD libraries to increase the complexity of reads ('spiking')
- Reduce loading concentrations of Illumina plates
- Potentially: filter out bad reads

# Quality scores across bases: RAD datasets

#### RAD1

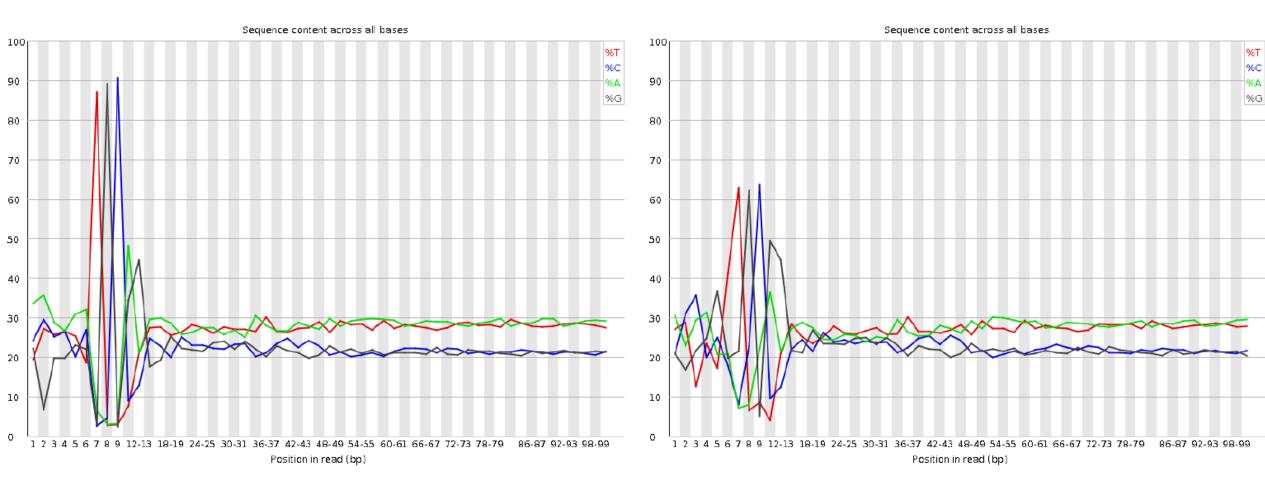
## Quality scores across all bases (Sanger / Illumina 1.9 encoding) This quality drop could be due to air bubbles passing through the flowcell or something else we will see later 1 2 3 4 5 6 7 8 9 12-13 18-19 24-25 30-31 36-37 42-43 48-49 54-55 60-61 66-67 72-73 78-79 86-87 92-93 98-99 Position in read (bp)

#### RAD2



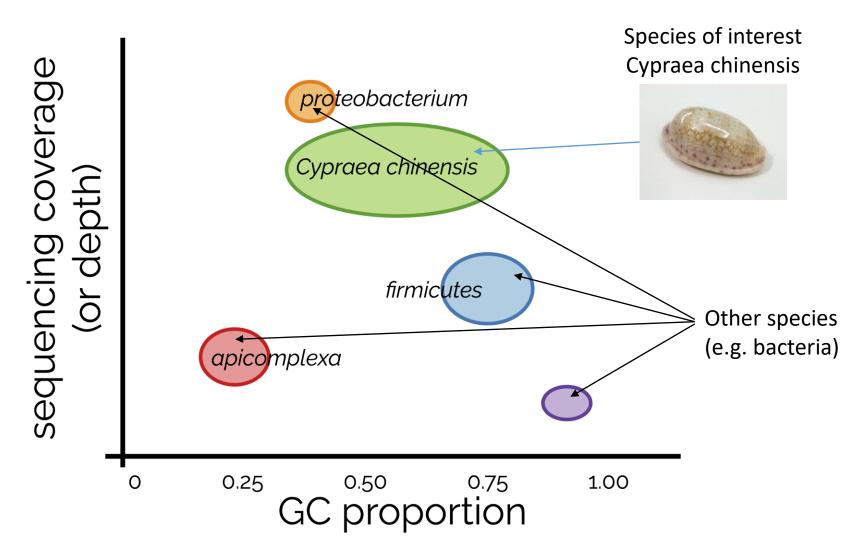
## Per base sequence content



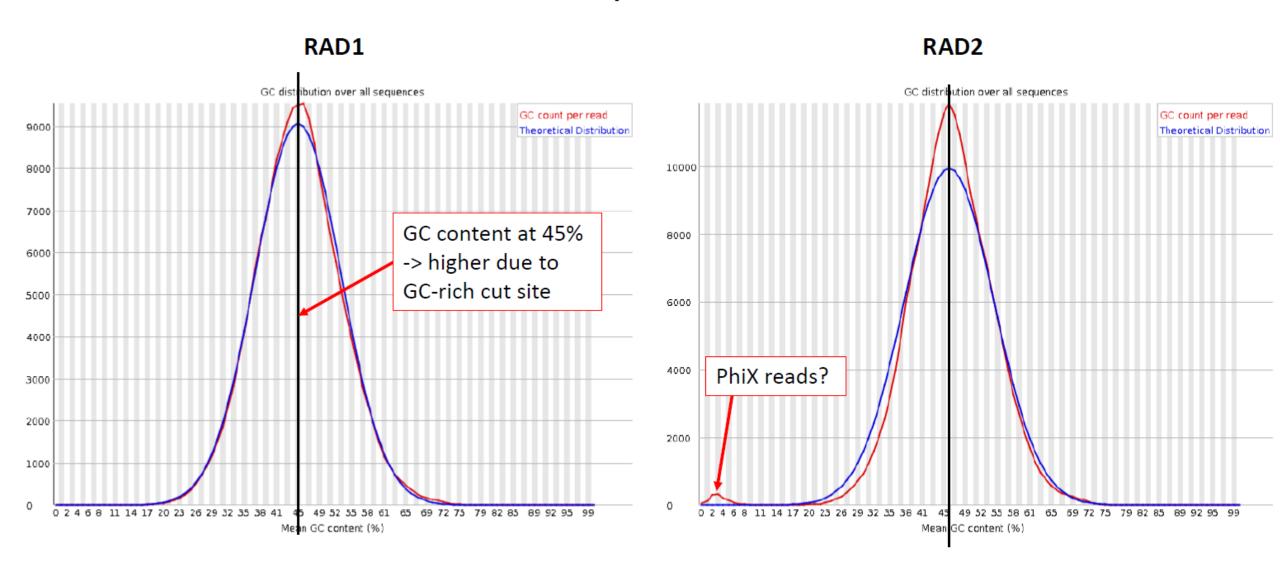


# Blobtools kit uses the GC versus AT proportion to identify contamination from bacteria or other organisms

(https://blobtoolkit.genomehubs.org/)



## GC distribution over all sequences



### **Examples of what you do not want to see:**

## Fastqc: Per sequence GC content

