ShortRead for quality assessment and data manipulation

Martin Morgan
Bioconductor / Fred Hutchinson Cancer Research Center
Seattle, WA, USA

8-10 June 2009

Running examples

'ChIP-seq'

- Solexa GA-II; 36mer single-end reads
- Whole-genome coverage
- Preliminary 'ELAND' alignment

'Pooled'

- Solexa GA-II; 36mer single-end reads
- Pooled sample, high coverage of three genes
- Searching for polymorphisms – SNP-like

'Barcode'

- Roche / 454 barcode; two zones
- Target length 200-300bp
- $\approx$ 20 bar codes (5' 8mers)

Input and Output

- Diverse input types, e.g., Solexa intensity, base call, alignment; MAQ text or binary, Bowtie; SOAP; fasta / fastq; tabular

```
> chip <- readAligned("./s_1_export.txt",
+     type = "SolexaExport")
> pool <- readAligend("./s_2.map", type = "MAQMap")
> bar <- read454("./454", ".*.fna$", ".*qual$")
```

- Ready access to data, leveraging standard R functionality

```
> reads <- sread(chip)
> qualities <- quality(chip)
> table(strand(chip))
```

- Output to fasta / fastq, tabular, genome browser tracks. . .

QA (quality assessment): reads per lane, Solexa GA-II

e.g., ‘chip’ data set

- Lane 5: internal control
- Typically 7-10M reads / lane
- 75-85% survive internal filtering, 50-65% align
- Lane 6: something amiss!

```
      read filtered aligned
1  8043779     0.75    0.62
2  8665770     0.77    0.66
3  7514774     0.80    0.68
4  8030556     0.79    0.68
5 11781447     0.72    0.84
6 11671931     0.59    0.21
7  8551614     0.77    0.65
8  8181482     0.76    0.63
```

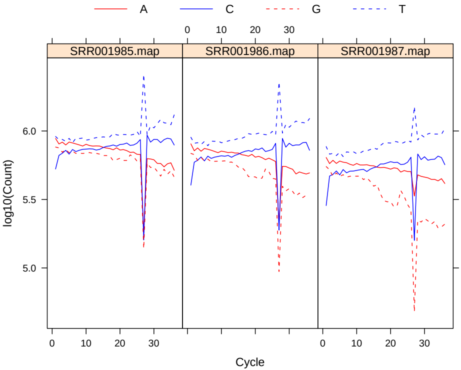
QA: base calls

- Uncalled nucleotides typically $< 1\%$
- Expected nucleotide frequency sample-dependent

```
     A    C    G    T      N
1 0.25 0.24 0.24 0.26 0.0150
2 0.26 0.25 0.25 0.24 0.0060
3 0.25 0.25 0.25 0.25 0.0061
4 0.25 0.25 0.26 0.23 0.0065
5 0.29 0.22 0.23 0.25 0.0062
6 0.24 0.29 0.27 0.19 0.0063
7 0.24 0.26 0.26 0.23 0.0070
8 0.24 0.27 0.27 0.22 0.0069
```

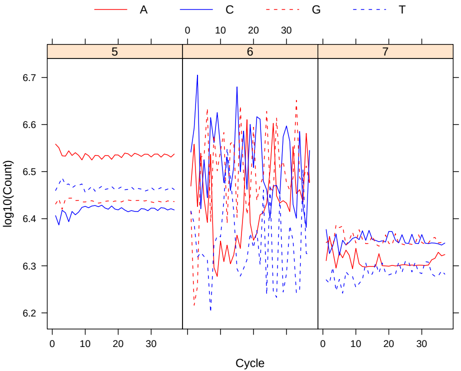
QA: reagent exhaustion and unusual base calls

- 3' exhaustion – directional trend in base call, e.g., due to reagent depletion; much less prevalent in GA-II
- Unusual base calls, e.g., due to machine malfunction
- Source: Chen et al., 2008, Cell 133: 1106-17. PMID: 18555785



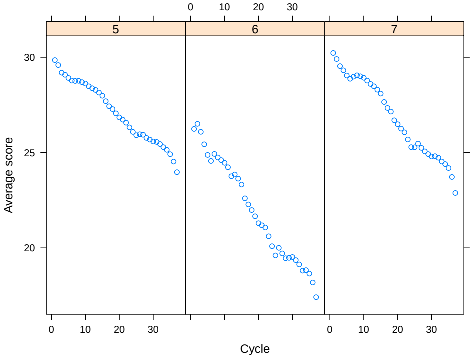
QA: alphabet-by-cycle synchronicity

- Lane 5: control; very consistent base calls
- Lane 6: reads dominated by relatively few sequences
- Lane 7: typical sample results; early synchronicity
- GA-I: first 1-2 bases show strong bias



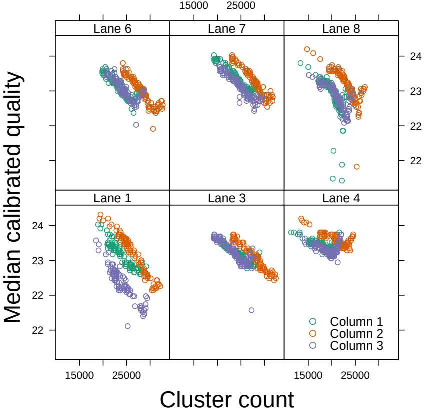
QA: tail quality

- Average base call quality (phred-like score) declines with cycle
- Sometimes abrupt changes (not illustrated)
- Often lane-specific, due to sample preparation and processing. Consequences for downstream analysis, e.g., ‘normalization’? processing



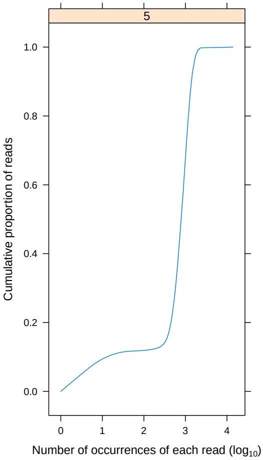
QA: quality / quantity trade-off

- Quality of base calls inversely related to quantity of reads



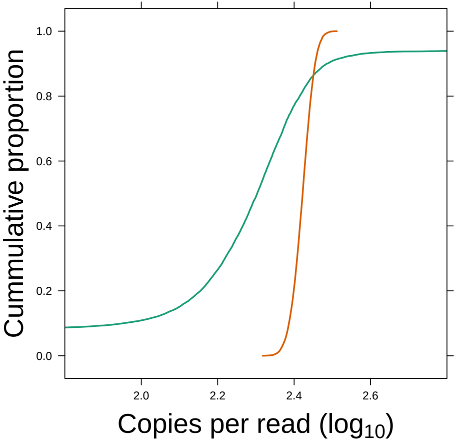
QA: frequent sequences

- Control lane, $\phi X174$ deep coverage
- Left: unique or nearly unique sequencing errors, 10-15%
- Right: highly repetitive, 5-10%



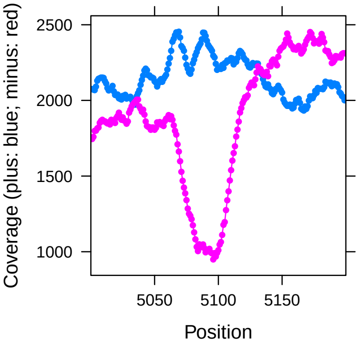
QA: frequent sequences

- Control lane, $\phi X174$ deep coverage
- Left: unique or nearly unique sequencing errors, 10-15%
- Right: highly repetitive, 5-10%
- Over-dispersion relative to uniform sampling: mappable genome, GC content, amplification bias, . . .



QA: alignment odditities

- pool: high coverage of small regions
- Close inspection: regions of unexpected low coverage. Single and double strand.
- Explanations: unmappable (e.g., repetitive sequence); primer similarity (filtered by upstream analysis); palindromes (failed sequencing PCR); poorly amplified (e.g., GC-rich)



ShortRead quality assessment report

- HTML quality assessment reports from diverse inputs
- Augments manufacturer reports
- Behind-the-scenes: the `qa` function distributes lane-level computations across MPI nodes, if available.

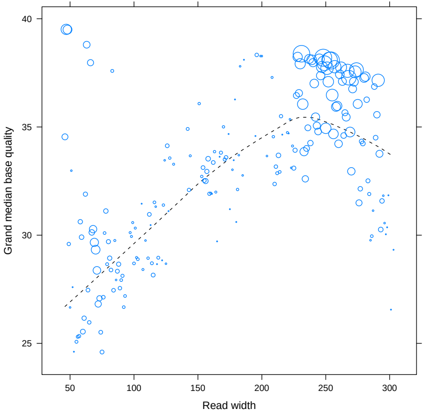
Examples

- Wang et al. Alternative isoform regulation in human tissue transcriptomes. Nature 2008 Nov 27;456(7221):470-6. PMID: 18978772
- `http://cbsresource.fhcrc.org/~mtmorgan/proj/GSE12946/qa_090502/`

```
> qa <- qa("./GSE12946", ".*.gz", type = "fasta")
> rpt <- report(qa)
> browseURL(rpt)
```

454 QA: read length / read quality

- ‘barcode’ data set, one zone
- Larger symbols indicate more reads
- Length and quality variation → quality gating



Common quality assessment issues

Illumina / Solexa

- Sample preparation artifacts, especially PCR prior to GA-II
- Base quality degradation, e.g., reagent exhaustion
- Read quality / quantity trade-off
- Nucleotide / dinucleotide bias?
- Sample-specific issues

Roche / 454 (preliminary)

- Terminal base quality
- Length heterogeneity
- Early indels