Short Reads

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

24 November, 2009

Short Reads

Biological questions.

- ChIP-seq; SNP discovery; RNA-seq; digital gene expression; de novo assembly.

Overall process – Illumina Genome Analyzer II.

1. Biological preparation, e.g., ChIP.
2. Library preparation: sonication, adapter ligation, size selection.
3. Cluster generation: bridge PCR, reverse strand removal.
4. Sequencing: florescent, reversibly terminated nucleotides.
5. Analysis.

GA II Read Characteristics and Throughput

Read characteristics

- 30-100bp.
- Single-end: one end of the amplified fragment.
- Paired-end: both ends of the amplified fragment, $\approx$ 200bp apart.
- Mate pair: larger genomic sequence, circularized, fragmented to span circularized location, paired end sequencing.

Throughput (24 November, 2009)

- Our runs: 80bp sequences, 20 million reads per lane, 8 lanes per cell.

Other technologies

- Roche / 454: 300-500bp reads, 1 million reads.
- ABI SOLiD: 60 gigabase, 1 billion reads / run. High-accuracy reads from 'color-space' model (no Bioconductor support for color space).
- Also: Helicos (single-molecule); PacBio; ...

Issues in Alignment and Experimental Design

Alignment.

- Numerous well-discussed issues related to read quality (deteriorates with cycle number), bias (GC-rich regions underrepresented), mappability (alignment algorithms avoid repeat regions), etc.

Experimental design.

- Illumina GA II 'Lane' as unit of sample replication. Important flow cell block effects, partly because technology moves very quickly.
- Multiplexing (several indivdiuals per lane) becoming increasingly important; likely barcode effects.
- Many studies do not include replicate samples, even though it seems obvious that this is required for down-stream quantitative analysis.

Bioconductor tools

- Biostrings (sequence representation; pattern matching); BSgenome, BSgenome.* (model organism whole-genome sequences).
- IRanges (ranged-based representations and manipulations).
- rtracklayer (track and genome browser interface), ShortRead (I/O and quality assessment);
- chipseq, ChIPseqR ChIPsim ChIPpeakAnno (ChIP-seq analysis); Genominator (RNA-seq); baySeq, DEGseq, edgeR (differential expression).
- HilbertVis (novel visualization).
- GenomicFeatures, biomaRt, org.*, . . . (annotation)
- . . .

I/O

ShortRead and Biostrings

- `readFASTA`, `read.DNAStringSet` sequence input.
- `readFastq` fastq sequence and quality scores.
- `readAligned` sequence, quality, and alignment information from a variety of aligners.
- `writeFASTA`, `writeFastq`

rtracklayer

- `import` and `export` browser track formats (bed, wig, etc.)

Examples of Sequence Manpipulation

Aligned reads (ShortRead)

- `tables` to summarize read occurences.
- `alphabetByCycle` to summarize nucleotide use per cycle.

Ranges and strings (IRanges, Biostrings)

- `alphabetFrequency` talies nucleotide use (also di- and tri-nucleotide variants, and sliding window calculations).
- `narrow`, . . . to reduce read width.
- Pattern matching, e.g., `trimLRpattern` for trimming left and right ends of reads; pairwise local and global alignment; whole-genome algnment with `matchPDict`.

Quality Assessment

```
> library(ShortRead)
> fls <- list.files("/path/to/folder", ".*map",
+     full = TRUE)
> qa <- qa(fls, type = "MAQMap")
> browseURL(report(qa, dest = tempfile()))
```

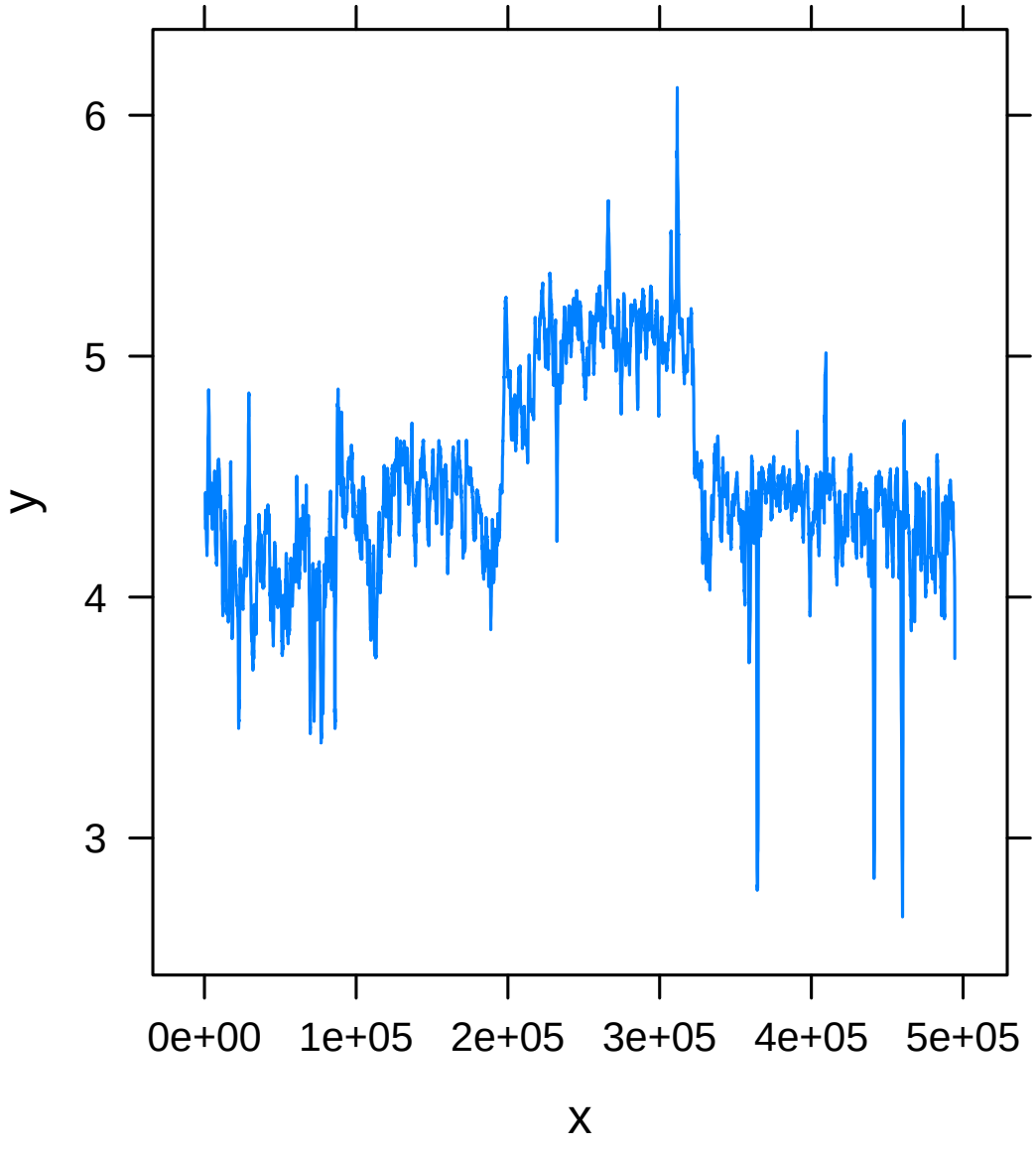
- qa summarizes contains QA summary information for subsequent computation.

A Forthcoming Development: Rsamtools

Storing aligned reads SAM/BAM.

- An indexed, random access, remote, slowly emerging standard.
- Rsamtools provides flexible access.

```
> source("../script/coverageplot.R")
> fl <- file.path("~/proj/a/1000g",
+     "NA19240.chrom6.SLX.maq.SRP000032.2009_07.bam")
> param <- ScanBamParam(what=c("pos", "width"),
+     which=RangesList(`6`=IRanges(0L, 500000L)),
+     flag=scanBamFlag(isUnmappedQuery=FALSE))
> bam <- scanBam(fl, param=param)
> cvg <- with(bam[[1]],
+     coverage(IRanges(pos, width=width), shift=-5000L))
> show(coverageplot(asinh(cvg)))
```



Asinh-transformed coverage from high-density 1000 genomes individual NA19240 Solexa sequencing on chromosome 6. There is a large copy number variant, and peaks of abnormally high and low coverage.

Summary

- Challenges for handling very large volumes of data.
- Role for R / Bioconductor in data exploratinn, quality assessment, non-standard alignment problems, down-stream analysis.
- Many exciting, unexplored questions – 'query' genomic regions for structural variants across many fully sequenced individuals; appropriate statistical modelling of base call and alignment errors; analysis of designed experiments. . .
- Area of very active development.