Computational Challenges in the Analysis of Short Read DNA Sequences

Martin Morgan (mtmorgan@fhcrc.org)
Fred Hutchinson Cancer Research Center
Seattle, WA, USA

October 5, 2009

Abstract

Short read DNA sequence data poses significant challenges for computational analysis. Here we survey and assess these challenges, providing creative solutions and possible directions for development. It is useful to distinguish between large scale public data such as the TCGA, 1000 genomes and ENCODE projects, and data generated with more modest resources. The size of primary data is a major computational hurdle. However, many analyses are most interesting after data has been reduced (e.g., by alignment to reference sequences) to manageable size. The computational challenges then involve formulation and design of appropriate statistical questions, domain-specific (e.g., ChIP-seq) analyses, integrative approaches that combine sequence and other data sources, and sequence-based annotation. These themes are illustrated with reference to several examples from our group.

Data

Sample preparation

- Whole-genome
- Enriched: e.g., transcription factor binding sites
- Focused: e.g., single contiguous genomic region

Short reads

- Simple: 'short' (35-150 bp) uniform length reads, e.g. Illumina; 10's of millions of reads
- Paired-end: non-sequenced insert (200-400 bp) between paired reads; 10's of millions of reads
- Intermediate: 150-300 bp variable length, e.g., Roche; 100's of thousands
- (ABI / SOLiD)

Domains

Applications

- Peak detection: transcription factors, methylation, histone modifications
- Relative abundance:
  - Digital gene expression
  - Splice variants
- Single nucleotide polymorphism
- (Assembly)

Experimental scope

- 'Lab' experiments, e.g, 10-20 flow cell lanes
- 'Mining' experiments, e.g., SRA
- 'Consortium' experiments, e.g., 1000 genomes

Challenges

Pre-processing

- Data volume
- Sample preparation and technology bias
- Quality assessment
- Normalization

Analysis

- Experimental design
- Statistical paradigms: counts and measurement error
- Applications: peaks, differential expression, splices, SNPs

Annotation

- Genome coordinates (e.g., ChIP-seq)
- Transcript centric (e.g., RNA-seq)
- Large reference resources

Pre-processing: data volume

Example: Illumina 'flow cell'

- Raw (images, intensities, base calls and quality measures): Terabytes
- Raw reads (reads and qualities, preliminary alignments): 100's of gigabytes
- Pre-processed (qa-filtered, aligned, 'normalized'): 10's of megabytes

Who carries the burden?

- Very large data: IT support (not our problem!)
- Raw reads: initial stages of analysis
  - Important for quality assessment, pre-processing stages
  - Strategies: streaming / batch processing; summary; sub-sampling
- After pre-processing: easily manageable – *good news!*

Pre-processing: sample preparation and technology bias

Wet-lab sample preparation

- PCR, ligation, contamination, ...

Technology

- Artifacts, e.g., leading base
- Amplification bias
- Error rates
- The mappable genome

Application-specific challenges

- E.g., miRNAs: short, incorporating primers

Pre-processing: quality assessment and data exploration

Basic characterization

- Read counts, nucleotide calls, base qualities
  - Often cycle-specific
- Manufacturer software vs. user exploration

Technology-specific features

- 454: variable length; high quality
- Unpaired vs. paired-end reads

Pre-processing: normalization

Sample preparation

- PCR artifacts
- *Within-sample variability*

Technology artifacts

- Amplification bias
- Limitations of historical (archived) data

Experimental design

- Blocking, e.g., Illumina flow cells; manufacturer reagent kits; often a significant temporal component
- Replication

Example: basic description

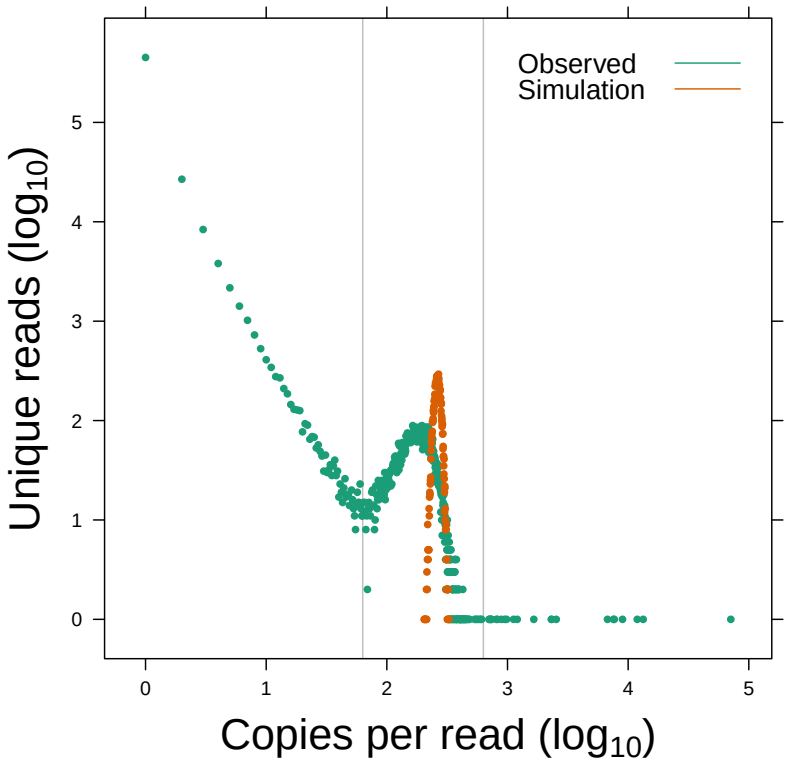
E.g., ChIP-seq quality assessment, Solexa GA-II

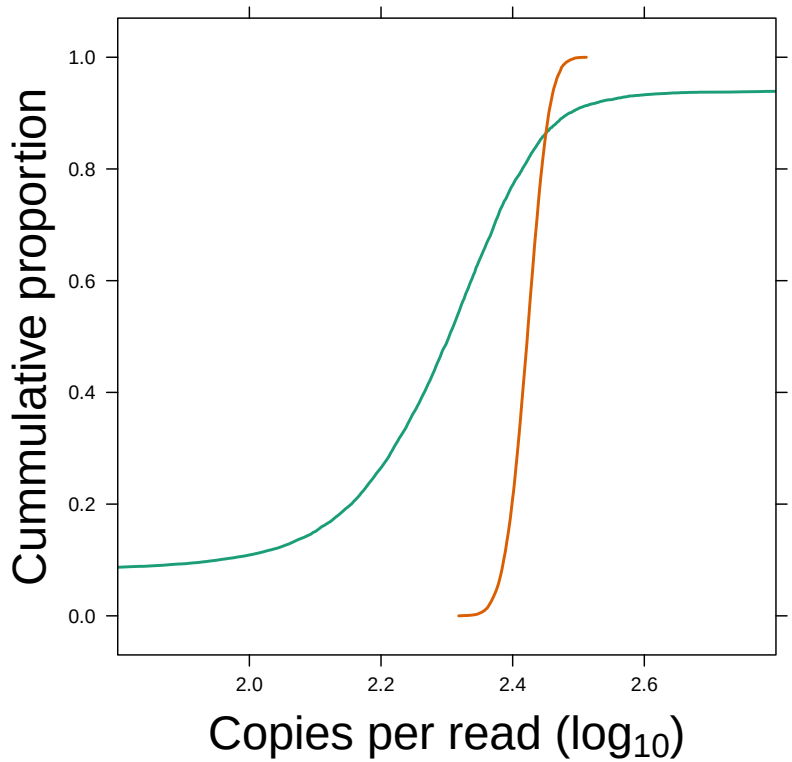
- Lane 5: internal control
- Typically 7-10M reads / lane
- 75-85% survive internal filtering, 50-65% align
- Lane 6: something amiss!

	read	filtered	aligned
1	8043779	0.752	0.620
2	8665770	0.774	0.655
3	7514774	0.800	0.676
4	8030556	0.791	0.675
5	11781447	0.717	0.844
6	11671931	0.590	0.206
7	8551614	0.769	0.645
8	8181482	0.761	0.630

Example: φX174 & systematic bias

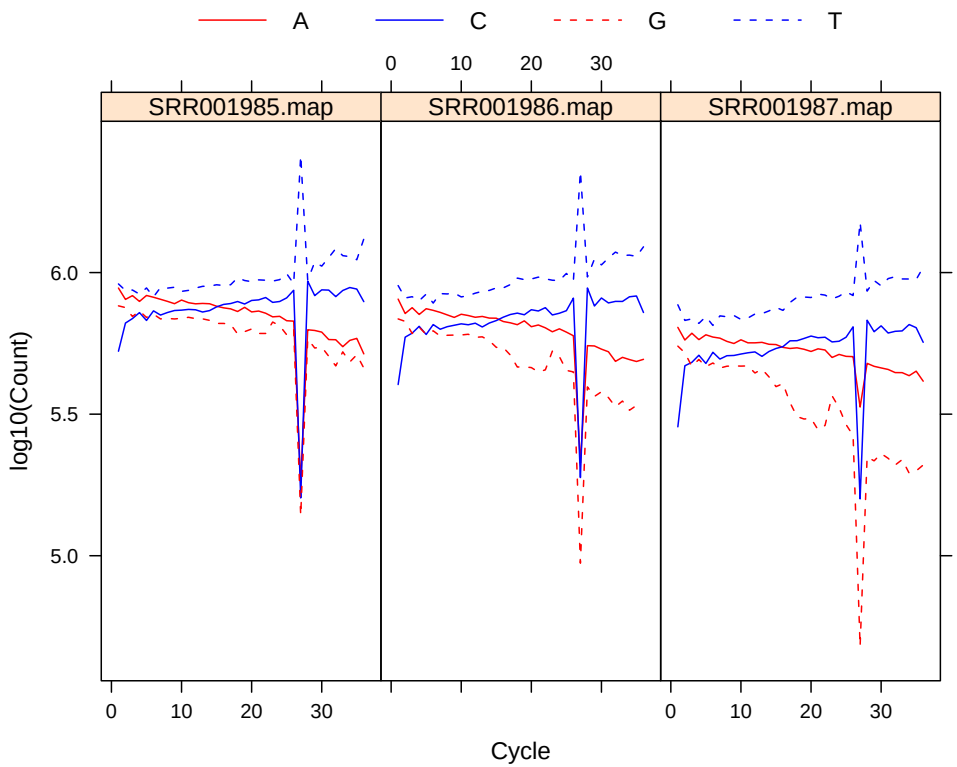
- Non-uniform coverage (amplification? sequencing bias?)
- Power-law error
- Adapter & primer sequence





Example: unusual base calls and end-drift in archives

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



Example: artifacts and base call errors

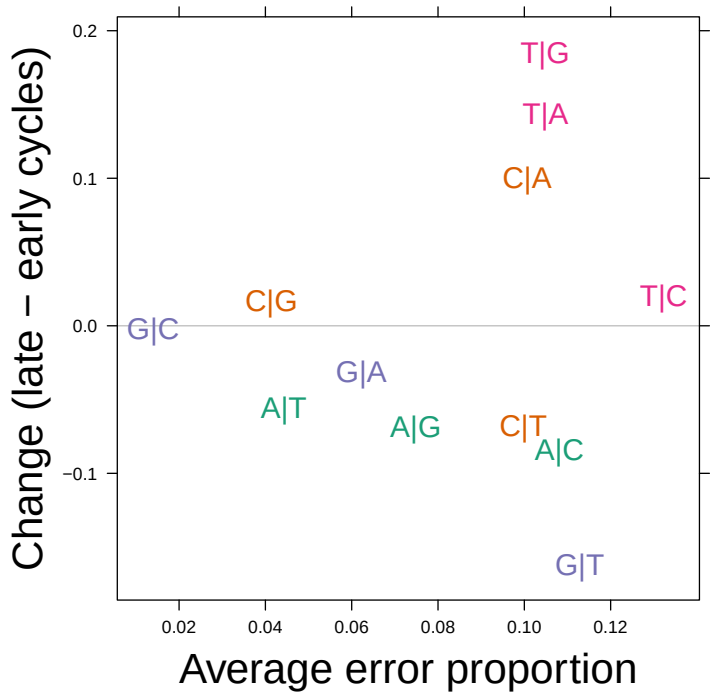
Technology artifacts

- Edit distance to Solexa adapters / primers

```
    0      1   ...      5
48863  34203   ...   8312
```

Sequencing error

- Base- and cycle-specific



Pre-processing: conclusion

Work flow

1. Manufacturer output; initial quality assessment
2. Alignment
3. Interactive exploration
4. Formal quality assessment / quality control
5. Normalization

End result

- High-quality alignment, largely independent of sequence or base quality information

Analysis: experimental design

'Lab' experiments

- Flow cell as natural experimental unit; strong batch effects
- Strong learning curve associated with adoption of new technology: later flow cells much better than earlier

'Mining' experiments

- Quality assessment; standardization

'Consortium' experiments

- Often similar issues, e.g., large-scale batch effects
- Biases induced by access restrictions, e.g., available only after patient death

Experimental design very important

- Avoid confounding treatment / batch effects
- Model batch effects associated with flow cell, run date

Analysis: statistical paradigms

Much like microarray data (!)

- Rectangular data; 'features' $\times$ 'samples'
- Easier to compare across samples than features (?)

Important application-specific issues, e.g., ChIP-seq

- Counts: distinct properties require appropriate error model
- Measurement 'features' discovered rather than *a priori*

Example: ChIP-seq work flow

Preprocess reads

- Duplicate reads as PCR artifacts?

Align to reference

- Mappable genome / multiply aligning reads

Identify peaks / islands

- Read extension (e.g., Kharchenko et al., 2008, Nature Biotechnology 26: 1351-9)
- Between-lane comparison, e.g,. pooling samples; control versus ChIP lanes

Coverage

- Number of (extended) reads aligned to each nucleotide

Islands

- Contiguous regions of non-zero coverage
- Characterize islands: area under the coverage curve, i.e., number of reads in the island

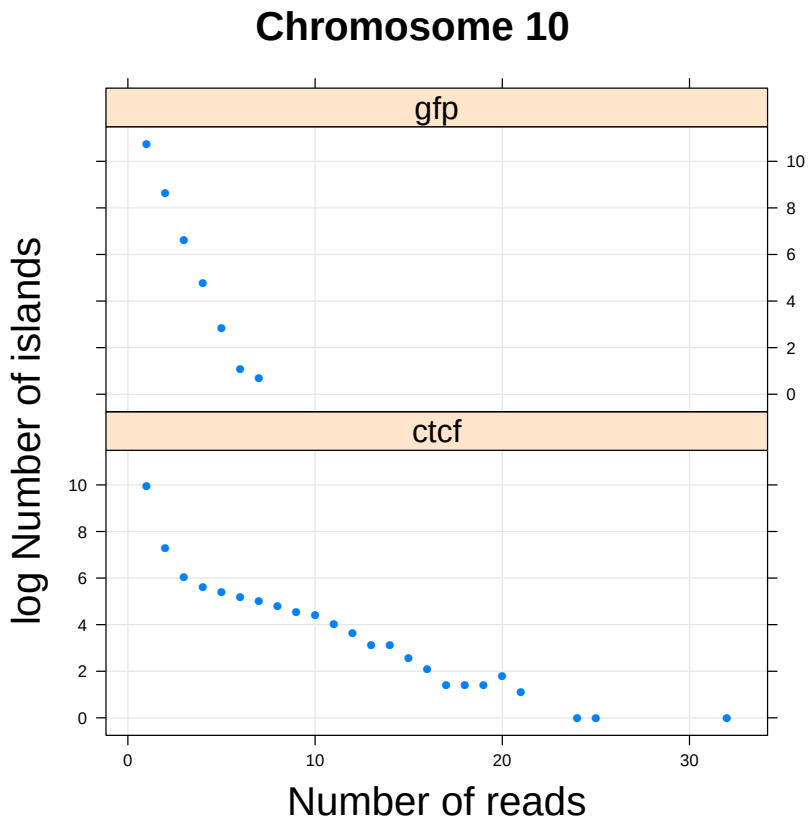
Example: ChIP-seq background versus signal

Null: $P(K = k) = p^{k-1}(1-p)$

- Random sample of reads from mappable genome
- Coverage K, with probability p that a read starts at a given position
- Estimate p by assuming islands of depth 1 or 2 derive from the null

Background threshold

- Usually strong evidence of departure from null
- Model-based and adaptive algorithms



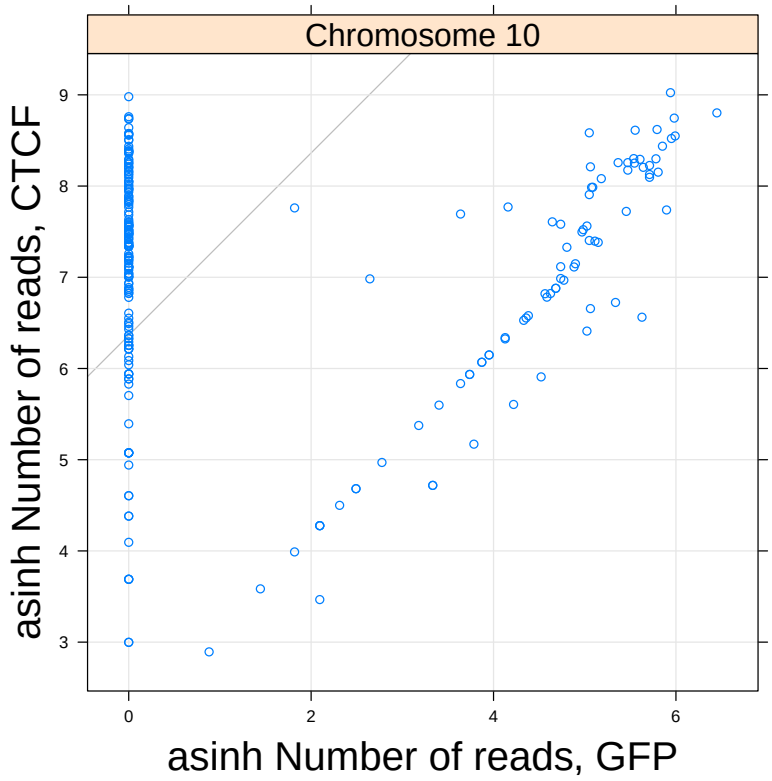
Example: ChIP-seq multiple lanes

Challenges

- Read number determines island statistics
- Lanes differ in read number
  - Sample prep. vs. biology

Possible solutions

- Down-sample to equal pool size, combine lanes, and identify islands
- Estimate scaling constant c from robust regression of $y = cx$.



Example: ChIP-seq as designed experiment

Summarized read counts

- Matrix with islands as rows, samples as columns; read counts are values

Statistical issues

- Islands are estimated, not defined *a priori*
- Data is count-based, not continuous; see Bioconductor edgeR for one solution

Example: experimental design

CAMDA 2009 ChIP-seq data

- Control ('Input') and ChIP ('Pol II') samples
- Flow cells likely a strong block effect

But. . .

- No assay with Input and Pol II in the same cell
- Some flow cells without replication

$\Rightarrow$ more efficient designs desirable

	SampleType	
FlowCell	Input	Pol II
FC12033	4	0
FC12044	1	0
FC12060	1	0
FC12170	4	0
FC12187	2	0
FC201WVA	0	5
FC20B5RA	0	2
FC4390	0	1
FC5817	0	3
FC6144	1	0

Analysis: conclusions

- Experimental design
- Appropriate statistical paradigms
- Application-specific issues

Annotation

Historically

- Gene-centric: 'top table' of expressed reporters; annotations; downstream analysis, e.g., GO
- Low-throughput: focus on one region of interest at a time
- Simple visualize, e.g., in the UCSC browser

Desirable

- Genomic coordinates, e.g., transcription factor binding sites
- Structured aggregations, e.g., transcripts
- Computable annotations
- Integrated visualization
- Large reference resources, e.g., 1000/1 genomes

Example: reference data base

Query 1000 genomes for common variants

- Specify ranges of interest, e.g., 5' promoters of all genes
- Query and summarize variation across 1000 genomes
- Do so interactively

Technical challenges

- Represent many genomes space-efficiently
- Perform range-based queries
- Meaningfully summarize ranged data

Example: Bioconductor approaches

rtracklayer

- Common track format I/O
- Browser navigation

Rsamtools (not yet released)

- BAM binary alignment format
- Selective (which and what) input of aligned reads
- Remote queries

But. . .

- Really want to query *across* 1000 genomes

Example: range-based queries

```
library(rtracklayer)
roi <- ## region(s) of interest
    RangesList('21'=IRanges(35500000, 35800000))
session <- browserSession()
snps <- track(session, 'snp130', roi) ## 2068 SNPs

library(Rsamtools)
archive <-
  'ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/pilot_data/
fl <- paste(archive, 'NA19240/alignment',
        'NA19240.chrom21.SLX.maq.SRP000032.2009_07.bam',
         sep='/')
p1 <- ScanBamParam(which=roi, simpleCigar=TRUE)
aln <- readAligned(fl, type='bam', param=p1) ## 290241 read
```

Directions

- Domain-specific applications, especially contributed by the user community
- Efficient memory management during pre-processing
- Range-based operations, including integration with external data sources
- Genomic coordinate annotations
- Multiple genome access and representation

Acknowledgments

- Robert Gentleman
- Patrick Aboyoun, Marc Carlson, Michael Lawrence, Hervé Pagès, Deepayan Sarkar, Zizhen Yao.
- Supported by award number P41HG004059 from the National Human Genome Research Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Genome Research Institute or the National Institutes of Health.