Using Bioconductor with high throughput sequence data

Overview

- this is very much an experiment
  - most modules are not complete
  - we need your help to design software that will address your needs
  - there is a large amount of infrastructure in place, but we need more use cases to design interfaces
- we won't spend too much time on:
  - matching of reads
  - worrying about quality scores
- we have tried to use cut down, but realistic, examples

Overview

- very large numbers of reads, typically short (< 100 nt for Solexa; <500 for 454)
- typically they have errors, some notion of quality of the base call etc
- they need to be mapped to a genome
- then once mapped we need to manipulate them; try to make some sense of the data
- questions will be experiment specific:
  - ChIP-seq: where are the peaks, what is under them
  - RNA-seq: what RNAs are we seeing; splice variants

Data issues

- the data are pretty large
  - both experimental and meta
- we will need tools that try to keep storage requirements to a minimum
- three basic strategies:
  - a pass-by-reference paradigm
  - views on large vectors, not subsets
  - compact representations

Glossary

- read: a nucleotide sequence provided by the technology (eg Solexa)
- island: a contiguous collection of reads mapped to a genome
- island count: number of reads in an island
- island depth: for a given locus the number of reads that overlap that locus
- summit: maximum height for an island
- peak: a subset of an island, with depth > k, for some k

Overview: Topics

- ShortRead:
  - reading in data, quality assessment
  - preprocessing
  - rely on the output of different manufacturers, and provide parsers

Packages: IRanges

- contains the basic infrastructure for external vector representations
- has the Ranges classes
  - important for all sorts of our tools
  - a Range is essentially two integer vectors (start and end) of an interval
- coverage: depth of coverage
  - use a run-length encoding to greatly reduce the size of the data

Genomes

- the BSgenome package contains infrastructure
- different genomes come in their own packages
- masks can be used to hide parts of genomes (repeat regions etc)
- SNPs and other ambiguities can be “injected” into the genome
- chromosomes are stored as external strings
  - read only

Biostrings

- string matching
- our own competitor to MAQ, SOAP and Bowtie: matchPDict
  - uses and Aho-Corasick methods
  - returns all matches
  - deals with indels
- currently contains lots of other functionality
  - suffix tree code
  - RNA-DNA-Protein translations
  - palindrome matching
  - alphabet frequencies (di-tri-and higher)
  - standard alignment methods

Visualizing

- we need to have methods to look at the data
  - it is large, complex and prone to errors
- some R level solutions
- rtracklayer can be used in two ways
  - produce tracks and push them to a genome browser
  - get data out of tracks that are downloaded from UCSC for example

Annotation

- typically we will care a bit about genomic context
- which peaks are in introns, exons, promoters, near genes, near microRNAs, near XXX
  - Bioconductor annotation packages
  - biomaRt
  - rtracklayer

Our Experiment

- you are going to get to use some real data, but since we have not published on it yet, we are not going to be able to give complete details
- you cannot use the data we are providing for anything after the course is over
- the data you will see come from a paired-end ChIP-seq experiment (can't tell you the transcription factor)

Experiment

- there are three lanes of data
  - the data are for two different mouse cell lines
  - two lanes have signal (a TF is active)
  - one lane is control (no TF)

Experiment

- the transcription factor is cross-linked to the DNA
- the DNA is sonicated
- an antibody to the TF is used to precipitate the TF + bound DNA
- cross-linking is reversed
- DNA collected and sequenced
- average fragment length is about 200nt
  - we did measure it

Experiment

- PCR is used to amplify the amount of DNA
  - so we tend to believe that any two reads that map to the same locus are probably due to PCR, not getting duplicate reads
- there are lots of artifacts
  - we see lots of islands of size 1
  - these seem to be some sort of background signal
- some visualization is going to be essential