Workshops

Next Generation Sequencing Workshop

March 10 - April 21 2010

Week 3 (March 31)

Functional genomic applications. Will include coverage of RNA-seq, ChIP-seq, and GRO-seq. Case studies will be emphasized.

There is an online discussion forum set up for this workshop. Workshop announcements will be posted there, and it is the best place to ask any workshop related questions, all teachers and organizers will be monitoring this forum closely. Workshop participants will need to register on forum website to obtain forum id before posting.

The speakers will be available to answer questions for this session on Friday April 2 from 1:00 to 2:00 PM in 102 Weill (small conference room).

Week 3 data files:
public directories ftp://cbsuftp.tc.cornell.edu/ngw2010/session3
protected directories ftp://usr@cbsuftp.tc.cornell.edu/ngw2010p/session3
NOTE: User id and password to access protected files have been e-mailed to registered participants. Replace "usr" in the link above with the user id you received.

Lecture 1. RNA-seq.
Speaker: Lalit Ponnala (Computational Biology Service Unit)

Lecture 1 slides.

1. brief description of the technology
2. types of biological problems that it can be used for
3. our case-study: maize transcriptome
4. (if time permits) methods to statistically analyze differential expression (simple test, more involved tests)

Working example

- dataset to use: maize Mo17
- software required: tophat + cufflinks
- input data files and formats set of commands to run (screenshots of running)
- output format, interpretation

Pitfalls

- spotty coverage (show browser screenshot)
- types of alternate splicing events: hard to clearly classify when seen on genome browser (especially AFE,ALE cases)
- ill-understood: add junction coverage while assessing transcript expression? (many methods do not seem to explicitly do this!)
- statistical issues: differential expression (q-value, p-value setting)

Exercise 1:
Download exercise instructions here.

- Run Tophat to align reads from two separate maize samples, named "zero" and "one" Run Cufflinks to calculate transcript and gene expression levels
- Run Cuffdiff to detect differential expression

Lecture 2. Chip-seq.
Speaker: Josh Waterfall (Lis Lab).

Lecture 2 slides.

Related experiments:

- Chromatin ImmunoPrecipitation (ChIP-seq)
- Micrococcal Nuclease (MNase-seq)
- DNase-I Hypersensitivity-seq
- Formalydehyde Assisted Isolation of Regulatory Elements (FAIRE-seq)
- Methylated DNA (also similar to SNP calling)
- Genome Run-On (GRO-seq)
- Oligo-capping

Advantages over array based methods:

- No cross-hybridization background
  - Unmappable portion of genome much smaller than repeat masked portion
  - Much more sensitive, basically limitless, lower limit of detection
  - More linear/quantitative dynamic range
- No restriction based on probe locations

Estimating amount of sequencing necessary:

- Amount of DNA needed for generating library.
- Re-sampling to determine peak calling saturation
- Bar-coding if less than one lane needed.

Most work is done with Illumina sequencing but work with Helicos, 454, and SOLiD have also been published.

All the applications we will cover assume there is a reference genome
After aligning reads to genome:
Identify enriched regions

1. Localized peaks (e.g. for sequence specific binding factors)
   a. Available tools: MACS, SPP, SISSRS, PeakFinder
   b. Structure between plus and minus strand reads
   c. Unmappable portions of genome
2. Large, extended bound regions
   a. No standard tools, several ad hoc approaches
3. Sources of background
   a. Non-specific (and possibly non-uniform) DNA pull-down
   b. Antibody cross-reactivity

Issues, Concerns, and Limits:

- Most current technologies (although this is changing) require library amplification and are thus sensitive to PCR bias.
- Current technologies still require large population of cells (it's hard to do biochemistry on a single cell).
- Most of these applications benefit more from increased read number rather than read length (once read length is more than ~ 25 or 30 bp).
- If your factor binds very repetitive DNA this can be problematic (but nearly every other assay is too).
- DNA ligase doesn't have significant sequence bias but T4 RNA ligase does (be careful of barcoding RNA samples before cDNA synthesis).
- Lack of standard methods for identifying extended enriched regions.

Exercise 2:
Download exercise instructions here.

- Using published ChIP-seq data (possibly human Stat1 before IFN treatment from Snyder lab)
- Run MACS with different parameter values
  - Look at differences in number of peaks called
  - In genome browser look at different peaks
- Maybe run another package (possible spp)
  - How many peaks were also called by MACS?
  - How many were not?
  - Look at them in browser and evaluate.
- Look at peaks in browser

Website credentials: login