Lectures 13: High throughput sequencing: Beyond the genome

Fall 2018
October 23, 2018
RNA-Seq uses 2nd gen sequencing machines to catalog full collection of RNA in a cell — known as the transcriptome.

When the reads are aligned back to the reference genome from the species of origin, a clear picture emerges.

This data can be used to identify both SNPs and expression, since the number of tags observed for each exon is proportional to the number of copies in the cell.

Omics

- Transcriptome - the set of all mRNAs present in a cell
- Proteome – proteins
- Metabolome/physiome - metabolites
- Microbiome – the collection of microbes present in an organism or other location
- Interactome

“In physics... the -on suffix has tended to signify an elementary particle: the photon, electron, proton, meson, etc., whereas -ome in biology has the opposite intellectual function, of directing attention to a holistic abstraction, an eventual goal...” From: ‘Ome Sweet ‘Omics. The Scientist 15(7), 2001
Omics

- Biologists have high-throughput methods for probing each -*ome*:
  - Transcriptome – RNA-Seq
  - Proteome – mass spectrometry, protein arrays
  - Microbiome – next generation sequencing
  - Interactome – yeast-two-hybrid
  - Regulome – ChIP-Seq

Lots of data for bioinformatics people to analyze!
RNA-seq: profiling the transcriptome

- **Technique:** sequence the total RNA produced by the cell

RNA-Seq uses 2nd gen sequencing machines to catalog full collection of RNA in a cell - known as the transcriptome.

When the reads are aligned back to the reference genome from the species of origin, a clear picture emerges.

This data can be used to identify both SNPs and expression, since the number of tags observed for each exon is proportional to the number of copies in the cell.
Read mapping

Sequenced reads can be aligned to the reference genome using an aligner, such as MAQ, bwa, Eland, Exonerate or Bowtie.

Aligners work as a black box to locate the most likely point of origin of each sequenced read.

The longer the reads, the more likely the aligner will find a unique (or best) point of origin. Most aligners do not require perfect matches.
Pile-ups

Pile-ups

Most reads fall into coding exons or UTRs
RNA-seq: profiling the transcriptome

- Technique: sequence the total RNA produced by the cell

- What is this good for?
RNA-seq: profiling the transcriptome

- Genome annotation (transcript assembly)
- Detect alternative splicing
- Obtain gene/transcript expression levels and detection of differential expression
- Allele-specific expression
- Small-RNA transcriptome (different protocol than regular RNA-seq)
All the uses of RNA-seq

Differential expression

The first step of the compare function is to identify locations with similar peaks in both sets of reads.

These points are plotted and a symmetrical best-fit regression line is calculated.

The distribution of all peak pairs around the line are calculated and points close to the line are removed.

Interesting

not interesting

Any points remaining are statistical outliers.

http://www.fejes.ca/labels/figures.html
RNA-seq protocol

1. Purify and capture mRNA from tissue or cells using beads

2. Fragment RNA by sonication or
2. Reverse Transcribe to cDNA

3. Reverse Transcribe to cDNA by random priming or
3. Fragment cDNA by sonication

4. Ligate sequencing adapters

5. Sequence and analyse millions of reads
Raw and Aligned Reads

- Raw data is a (large) set of sequences
- Typical file format is FASTQ
  
  ```
  @HWI-EAS255_4_FC2010Y_1_43_110_790
  TTAATCTACAGAATAGATAGCTAGCATATATTT
  +
  hhhhhhhhhhhhhhhhhhhhhhhhhhhhdRehdh
  ```

  Read identifier  Bases called  Base quality codes

- Alignment to genome is done by efficient indexing
- Aligned reads in SAM format
  
  ```
  @HWI-... 163 chr19 9900 10000 16M2I25M
  ```

  Read identifier  Where this read matched  Start and end positions  Codes for match: 16 matches, 2 extra,...
Cataloging the transcriptome

• Transcriptomics involves studying expression at
  – Spatial resolution: tissues, individuals, location
  – Temporal resolution: circadian, seasonal, lifetime
Inter-Genic Reads

• Many reads reflect unannotated genes: opportunity to discover new genes
RPKM – A Simple Normalization

• Different numbers of counts per sample (sequencing depth)
• Divide counts in a region of interest (a genomic region or a gene or an exon) by all counts (reads per million reads -RPM)
• Genes have different lengths: divide also by length of gene
• Obtain RPKM (reads per kilobase of exon per million reads)
  – Some use FPKM (fragments/kb/Mr)
ChIP-seq uses chromatin immunoprecipitation and massively parallel sequencing to locate genome-wide protein-DNA binding events.

Proteins touching DNA are fixed in place with a cross-linking agent.

Cross-links are broken and only DNA fragments from binding sites remain.

DNA is fragmented and complexes are harvested with targeted antibodies.

They can then be sent for sequencing.

http://www.fejes.ca/labels/Chip-Seq.html
Comments on ChIP-seq

• Genome-wide mapping of transcription factor binding sites

• Computational problems:
  – Peak calling
  – Still need motif finders, but makes the problem easier
Variants

• Apply the methodology to RNA: map RNA-binding sites in mRNA that interact with specific RNA-binding proteins
• CLIP-Seq (cross-linking immunoprecipitation sequencing)
• RIP-Seq (RNA immunoprecipitation sequencing)
Other sequencing-based techniques

- Methyl-seq, BS-seq: methylation
- Chromosome conformation capture (3C-4C-5C-HiC): spatial organization of chromosomes

http://en.wikipedia.org/wiki/Chromosome_conformation_capture
Other sequencing-based techniques

- Methyl-seq, BS-seq: methylation
- Chromosome conformation capture (3C-4C-5C): spatial organization of chromosomes
- seqFold: RNA secondary structure
- DNAase-seq
- And many more!

http://en.wikipedia.org/wiki/Chromosome_conformation_capture
Read mapping

All the sequencing-based techniques require read mapping as a first step.

Existing alignment tools are not fast enough → need new algorithms!
Read mapping

• How is the problem of read mapping different than sequence alignment as we have considered it until now?