RNAseq / ChipSeq / Methylseq
and personalized genomics

7711 Lecture

Subhajyoti De, PhD

Division of Biomedical Informatics and Personalized Biomedicine, Department of Medicine
University of Colorado School of Medicine

September 11, 2014
RNA sequencing
- The biology of RNA: types of RNA, intron-exon structure, alternative splicing
- Microarrays to quantify RNAs
- RNAseq protocols
- Quantification of RNA abundance
- Quality control in RNAseq data
- Microarrays vs RNAseq
- Software and their basic principles
- RNA editing

Chipseq / methylseq
- The biology of chromatin – histone modifications, CpG methylation, nucleosome
- Quantification of abundance of chromatin signature
- Software and their basic principles
- Combination of chromatin marks

Variant detection
- Germ line and somatic mutations
- Quality control issues
- Software and their basic principles

Genomics and Personalized Medicine
RNA sequencing
Types of RNA:
(a) mRNAs: codes for protein/peptide sequences
(b) tRNAs: small RNA, necessary component of protein translation
(c) microRNAs: small RNA that plays a regulatory role in major biological processes; typically suppress expression of target genes.
(d) Ribosomal RNA: small RNAs that are key components of ribosome
(e) Long non-coding RNAs: Large multi-exonic RNAs. Emerging reports link them to major biological processes
(f) Other non-coding RNAs: such as circular RNA

Kapranov, BMC Biology, 2010
1. Quantile normalization
2. RMA normalization
3. GCRMA
4. Mas5
5. ...

Lim et al. Bioinformatics. 2007
RNA sequencing > RNAseq protocol

Removal of Ribosomal and other types of RNAs. PolyA selection or Ribo-minus.

Sequencing using:
- Pyrosequencing (454 Technologies)
- Solexa sequencing (Illumina HiSeq)
- Sequencing by ligation (SOLiD)
- Ion Torrent semiconductor sequencing
- Nanopore sequencing
- Single molecule real time sequencing (PacBio)

Read length: 50bp – 10,000bp
Sequencing volume: ~ million reads/sec
Sequencing accuracy: 90% - 99.99%
RNA sequencing > Quantification of RNA abundance

RNAseq has excellent reproducibility

RNAseq has excellent dynamic range

RPKM (Reads per kilo base per million) is a measure of expression level of a genomic entity.

\[
\text{RPKM} = \frac{\text{transcription\_reads}}{\text{transcription\_length} \times \text{total\_assembly\_reads\_in\_run}} \times 10^9
\]

\[
\text{RPKM} = 10^9 \times \frac{C}{N \times L}
\]


Paired end read: FPKM (fragments per kilobase of exon per million fragments mapped)
RNA sequencing > Quantification of RNA abundance

RNA sequencing > Quality control in RNAseq data; Microarrays vs RNAseq

**a**

RNAseq has excellent reproducibility i.e. low technical variation


**b**

There is considerable amount of biological variation


**c**

The technical variation typically follows Poisson distribution

Mortazavi et al. Genome Res. 2008

There is robust concordance between microarray and RNAseq data for the same sample.

The GC-bias is not always straightforward.

Hansen et al. Biostatistics. 2011
RNA sequencing > Software

**a** Map paired cDNA fragment sequences to genome

**b** Assembly

**c** Minimum path cover

**d** Abundance estimation

**e** Maximum likelihood abundances

**Steps 1**
- Condition A
  - Reads
  - TopHat
  - Mapped reads

**Steps 2**
- Condition B
  - Reads
  - TopHat
  - Mapped reads

**Steps 3–4**
- Assembled transcripts
  - Cufflinks
  - Assembled transcripts
  - Cuffmerge
  - Final transcriptome assembly

**Steps 5**
- Differential expression results
  - Cuffdiff
  - Mapped reads

**Steps 6–18**
- Transcripts and their abundances
  - CummeRbund
  - Expression plots
**RNA sequencing > Software**

- **Integrative Genome Viewer / Broad Institute**
- **Evaluation**
  - Cuffdiff
  - DESeq
  - edgeR
  - limma
  - Voom
  - PoissonSeq
  - beySeq

| Evaluation                                                                 | Cuffdiff | DESeq      | edgeR      | limma
|                                                                           | Voom     | PoissonSeq | beySeq     | Voom |
|                                                                           |          |            |            |          |
| Normalization and clustering                                               | all methods performed equally well |
| DE detection accuracy measured by AUC at increasing qRT-PCR cutoff         | decreasing | consistent | consistent | decreasing |
|                                                                           | increases up to log expression change ≤ 2.0 | consistent |
| Null model Type-I error                                                    | high number of FP | low number of FP | low number of FP | low number of FP |
|                                                                           | low number of FP | low number of FP | low number of FP | low number of FP |
| Signal-to-noise vs. p-value correlation for genes detected in one condition | poor | poor | poor | good |
|                                                                           | moderate | good |
| Support for multi-factoried experiments                                    | no | yes | yes | yes |
|                                                                           | no | yes | yes | no |
| Support DE detection without replicated samples                            | yes | yes | yes | no |
|                                                                           | yes | yes | yes | no |
| Detection of differential isoforms                                          | yes | no | no | no |
|                                                                           | yes | no | no | no |
| Runtime for experiments with 3-5 replicates on 12 dual-core 3.33GHz, 100G RAM server | hours | minutes | minutes | minutes |
|                                                                           | minutes | seconds | hours | hours |

Rapaport et al. Genome Biol. 2013
### RNA sequencing > RNA editing

**A to I RNA editing modifies double-stranded RNA**

![Diagram showing A to I RNA editing](Diagram.png)

**Widespread RNA and DNA Sequence Differences in the Human Transcriptome**

*Li et al. Science. 2011*

![Diagram showing RNA and DNA sequence differences](Diagram.png)
How do you know if the RNAseq data looks ok?

S1: Untreated sample: 6 million reads
S2: Treated sample: 4 million reads

Depth of coverage plot

Differential expression MA plot
Chipseq / methylseq
Chipseq / methylseq > Chromatin biology– CpG methylation, histone modifications, nucleosome
Chipseq / methylseq > Quantification of abundance of chromatin signature

<table>
<thead>
<tr>
<th>Enrichment</th>
<th>MEDIP-Seq</th>
<th>MBD-Seq</th>
<th>RRBS</th>
<th>MethylC-Seq</th>
<th>Infinium array</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anti-5mC</td>
<td>MBD</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
</tbody>
</table>

Quantification of methylation

- Sequenced reads mapped to genome.
- Peaks of enriched methylation identified.

Bisulfite-converted sequence reads mapped to genome. Distinction of bisulfite-converted bases (T) and preserved cytosines (C) when compared to reference genome reveals positions of methylation (shaded blue).

Normalised ratio of hybridisation to methylated vs unmethylated probes on arrays results in a calculated percentage methylation ($\beta$ value).

Genomic DNA isolation
- MspI digestion
- Filling in and A-tailing
- Methylated Illumina adapter ligation
- Gel size selection
- Bisulfite conversion
- PCR amplification
- Final RRBS library QC

Chipseq / methylseq > Quantification of abundance of chromatin signature

Isolate genomic DNA

Cross-link whole cells with formaldehyde

Sonicate DNA to produce sheared, soluble chromatin

Add protein-specific antibody

Immunoprecipitate and purify immunocomplexes

Reverse cross-links, purify DNA and prepare for sequencing


Isolation of cells of the immune response
- Use the correct number of cells: $1 \times 10^6$ to $10 \times 10^6$
- Collect biological replicates of cells
- Choose an appropriate control for antibody specificity (knockout or RNP knockdown)

Fragmentation by sonication or MNase treatment
- Shear chromatin to a size range of ~150–300 bp
- Sonicate chromatin extracts for non-histone proteins
  * Sonication conditions should be determined empirically for each cell type
- Treat chromatin extracts with MNase for analysis of histone modifications
  * Do not overdigest chromatin

ChIP analysis of histone modifications, transcription factors or epigenetic regulators
- Select antibody: monoclonal versus polyclonal
- Choose reference control (Input or IgG)
- Perform ChIP with established protocols
- Purify DNA

Library construction
- Do end repair and adapter ligation
- Perform PCR using primers compatible with sequencing platform
  * Avoid overamplifying DNA

Sequencing
- Determine sequencing depth on the basis of the prevalence of binding throughout the genome; more sequencing tags may be needed for diffuse signals (such as H3K27me3)
- Perform single-end or paired-end sequencing

Zhao. Nature Immunology. 2011

Chipseq / methylseq > Software and their basic principles

Sequence alignment
(no longer a bottleneck for data analysis)

Inspect data quality
• Check report from vendor-supplied analytical pipeline
• Visually inspect with a genome browser
• Identify motif in tag-enriched region
• Examine the profile at certain genomic features
• Confirm by quantitative PCR

Peak calling (choose the right tool)
<table>
<thead>
<tr>
<th>Type of peak</th>
<th>Example</th>
<th>Representative tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Broad</td>
<td>H3K27me3</td>
<td>CCAT, SICER</td>
</tr>
<tr>
<td>Sharp</td>
<td>CTCF</td>
<td>MACS</td>
</tr>
<tr>
<td>Sharp &amp; broad</td>
<td>Pol II</td>
<td>ZINBA</td>
</tr>
</tbody>
</table>

Zhao. Nature Immunology. 2011
These programs produce very different peaks in terms of peak size, number, and position relative to genes.


Table entries show the overlap of peaks identified by each program with those identified by other programs. Each column entry gives the percent of base pairs belonging to peaks identified by the specified program that were also identified as in peaks by other programs. Peak height gives the mean difference between the highest and lowest read counts of the peaks identified by that program.

doi:10.1371/journal.pone.0025260.t002
Chipseq / methylseq > Combination of chromatin marks

Chromatin marks are associated with one another

Chrom-HMM

How do you compare the peaks?
Variant detection
(recap from previous lectures...)
Haemophilia
Type-II diabetes
Early onset obesity

Neurofibromatosis type 1 (NF1)
Cancer

Somatic mutations are common in healthy human tissues

De. Trends Genet. 2011
Variant detection > Quality control issues

Variant calling overview

- Input reads (BAM or fastq)
  - Alignment
  - Base quality recalibration input
  - Aligned reads (BAM)
  - Callable regions (BED)
    - Parallel by callable regions
      - chr1:1000-2000
        - Prepared reads in region (BAM)
      - chr1:5000-6000
        - Prepared reads in region (BAM)
        - Post-alignment preparation: recalibration and realignment

- Variant calling: supports multiple approaches and multisample calling
  - Called variants (VCF)
    - Merge variants
      - Combined raw variants (VCF)
        - Filter variants
          - Final annotated variants (VCF)
            - Variant evaluation report
              - Variant database (GEMINI)

An example: GATK pipeline
(Probably covered in previous lecture...)

McKenna et al. Genome Res. 2011

- Processed BAM file
  - Realigned BAM file
    - BWA align
      - Sample Bam File
        - Validate post-process BAM
          - Validate pre-process BAM
            - Mark Duplicates
              - Realigner Target Creator
                - Indel Realigner
                  - Count Covariates
                    - Validate Covariates
                      - Table Recalibration
                        - Analyze Covariates
                          - Count Covariates
Variant detection > Software and their basic principles

There is reasonable overlap among the mutation calling software

O’Rawe et al. Genome Medicine. 2013
Detection of variants from RNAseq / Chipseq / methylseq data

- RNAseq:: GATK RNAseq variant calling workflow for calling mutation from RNAseq data

- Methylseq:: BisSNP for detecting SNP from methylseq

- ChIPseq:: Simultaneous SNP identification and assessment of allele-specific bias from ChIP-seq data (Ni et al. BMC Genomics. 2012)

- Most of the normal software with suitable improvisation.

Detection of low frequency mutations in heterogeneous samples

- Mutect
- Mutsig
- JointSNVmix
- Many more...
Genomics and personalized Medicine
Systematic discovery from clinical samples

**ARTICLE**

**Comprehensive genomic characterization of squamous cell lung cancers**

Long squamous cell carcinomas is a common type of lung cancer, causing approximately 460,000 deaths per year worldwide. Genetic alterations in squamous cell lung cancers have not been comprehensively characterized, and no molecularly targeted agents have been specifically developed for its treatment. As part of The Cancer Genome Atlas, here we profile 178 lung squamous cell carcinomas to provide a comprehensive landscape of genomic and epigenetic alterations. We show that the tumour type is characterized by complex genomic alterations, with a means of 360 exonic mutations, 176 genomic rearrangements, and 1222 amplification events per tumour. We find mutually exclusive and recurrent activating mutations in several pathways, and genome-wide increases in DNA methylation. Finally, 32% of tumours harbor focal loss-of-function mutations seen in the RELA-A19F2 mosaic subset. Significantly altered pathways included NOTCH1 and KRAS2. 34% of tumours harboured RELA-A19F2 mosaic subset. We identified a potential therapeutic target in most tumours, offering new avenues of investigation for the treatment of squamous cell lung cancers.


**Application of genome sequencing to detect disease progression**

**LETTER**

**Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing**

We describe the results of whole-genome sequencing of primary and relapse samples from 8 patients with relapsed acute myeloid leukemia (AML). The most commonly mutated genes were TP53, ATM, CDKN2A, and ARID1A, which were also found in primary cases. Mutations in TP53 were most often observed in relapse samples from patients who had received chemotherapy. We also identified novel cancer gene mutations in a subset of patients. The clonal evolution of the disease was as a result of chemotherapy.


**Genome sequencing shaping diagnosis and treatment**

**Acute promyelocytic leukemia**

One patient

7 weeks for genome sequencing and analysis

Actionable gene fusion detected

Changed the treatment plan for the patient

Welch et al, JAMA, 2011
Subhajyoti De, PhD
Assistant Professor
Department of Medicine, University of Colorado School of Medicine
Anschutz Medical Campus

Email: subhajyoti.de@ucdenver.edu
Phone: +1 303 724 6461

Webpage: http://www.sjdlab.org/