Applications of RNA -seq in Microbiology

STUDENT: CHEUNG YUK YAM, ANDY

SUPERVISOR: PROF MAMIE HUI

JOINT GRADUATE SEMINAR DEC 2016

DEPARTMENT OF MICROBIOLOGY,

FACULTY OF MEDICINE,

THE CHINESE UNIVERSITY OF HONG KONG

Content

1. What is RNA-seq?
2. Advantages over microarray- and tag-based approaches
3. Application: RNA-seq of pathogens
4. Application: RNA-seq of host
5. Application: Dual RNA-seq?

   - progress

   - major challenges
6. Conclusion

What is RNA-seq?

- massively parallel sequencing of RNA
  (in fact, the corresponding cDNA)
- Typically, all RNA molecules -> reverse transcribed into cDNA
  -> (amplification based sequencing or single-molecule sequencing)
  -> deep sequencing by next-generation sequencing platforms
- After sequencing, the obtained sequence reads are mapped onto a reference genome to deduce the transcriptome

Next generation sequencing

	Roche/454	Illumina	Ion Torrent	Pacific Biosciences
Detection	Light (by luciferase)	different base different fluorescent (1 base a time)	Proton (pH)	fluorescent (No need to halt between read steps)
Read length	~1kb	~150bp	~200-400bp	over 10 kb
Max no. of read	~2 million	6 billion	>6 million	10^4
Remarks	Fragment ends -> adaptors ->beads, (1fragment per bead) PCR needed	fragments -> adaptors -> slide PCR needed	Fragment ends -> adaptors ->beads, emulsion PCR	SMS

Timeline of RNA-seq

- In 2008, 1st genome-wide RNA-seq experiments
  - one in mice and several in humans
  - yielded ~ 5–15 million reads per lane (Sultan et al. 2008 & Marioni et al. 2008)

By pooled reads,

  - ~30 million–100 million reads
  - Read lengths: 25–32bp (short)

- Later, RNA-seq was applied to numerous pathogens e.g. Salmonella, Listeria, Helicobacter, etc) (Perkins et al. 2009, Oliver et al. 2009 & Sharma et al. 2010)
  - ~5 million reads per sample
  - increased read lengths: ~36–40 bp

Timeline of RNA-seq

- Nowadays,
  - >6 billion reads of >150 bp in a single run;
- However, upper limits of sequencing resolution: not yet been reached
- ∵ advancement in third generation sequencing

  ∴ both read number and read length: expected to increase

Advantages over microarray- and tag-based approaches

- -theoretically infinite dynamic range
- linear dynamic range for RNA-seq
  - >4 orders of magnitude
  - now approaching 6 orders of magnitude
  - comparable to gene expression changes in eukaryotic cells
- No prior knowledge of RNA sequences needed
- can identify novel transcripts
  - e.g. a single RNA-seq study of mouse myoblasts (Trapnell et al. 2010)
    - identified ~4,000 previously unknown transcripts

Advantages over microarray- and tag-based approaches

- improve the annotation of the genomes
  - provide extensive information on transcription start sites
  - location of the 5‘ and 3‘ UTRs of known genes
  - report new ORFs
- development of strand specific RNA-seq
  -preserves information about the directionality of a transcript
- noncoding and antisense transcripts
- characterization of operons in bacteria

RNA-seq of pathogens

- Prokaryotic pathogens

  e.g. Salmonella, Listeria, Helicobacter, etc

  (Perkins et al. 2009, Oliver et al. 2009 & Sharma et al. 2010)

- Eukaryotic pathogens

  e.g. *Candida albicans*, *Plasmodium falciparum*, etc

  (Bruno et al. 2010 & Sorber et al. 2011)

RNA-seq study of *Helicobacter pylori* in 2010

- selective pretreatment of RNA:

  -exonuclease: degrades processed RNAs (rRNAs and tRNAs)

  leaves mRNAs and small non-coding undigested

- RNA analyses: exponential phase vs during acid stress

  (which the bacterium normally experiences in the stomach)

- identified 1,900 transcription start sites
- grouped ~1,700 protein-coding genes into 337 operons
- urease (*ure*) operon (virulence; upregulated under acidic stress)
- express a lot of noncoding RNAs, including antisense RNAs to 46% of all genes

(Sharmaet al. 2010)

In vivo RNA-seq vs *in vitro* RNA-seq

- environment in an animal host: differ from *in vitro* models
  - gene expression differs?
- *Vibrio cholerae*

  -caecum of infected rabbits or intestine of infected mice

  -> RNA-seq:

  identified 39 transcripts (out of 478): altered expression

  in both animal models vs. *in vitro* culture
- well-characterized virulence factors

  e.g. cholera toxin and the type IV pilus TCP

(Mandlik et al. 2011)

RNA-seq of host transcriptomes

- increasing studies: response of mammalian host cell to infection

Pathogen	Host cell(s)	Gene upregulated	reference
Schmallenberg virus	bovine cells	antiviral genes	(Blomströmet al. 2015)
West Nile Virus	macrophages	interferon related genes	(Qian et al. 2013)
Salmonella enterica	macrophages & HeLa cells	cytokine IL-6 and IL-10 mRNAs	(Schulte et al. 2011)
Candida albicans	human endothelial cells & oral epithelial cells	signaling pathway protein	(Liu et al. 2015)

- Dual RNA-seq:
  - revolutionize the study of host–pathogen interactions
  - new molecular insights in pathogenesis and immune response

Dual RNA-seq: major challenges

- Different RNA contents
- eukaryotic cells: 10–20 pg of total RNA
- bacterial cells: ~0.1pg
  - ~100–200 times difference
- In practice, a single infected host cell will contain multiple bacteria
  - smaller difference
  - ~10–20-fold in most infection models

Dual RNA-seq: major challenges

- heterogeneity of the RNA
- Eukaryotes: miRNAs, long non-coding RNAs, small nuclear RNAs and small nucleolar RNAs
- Bacteria: small non-coding RNAs
- E.g. eukaryotic mRNAs: a poly(A) tail: increase mRNA stability.

  bacterial mRNAs: rarely contain a poly(A) tail: as a tag for degradation

  -> selection by Oligo (dT) Primer

  -> stable transcripts (eukaryotic host) versus

  transcripts undergoing degradation (bacterial pathogen)
- ∴ enrichment for specific RNAs is not recommended

Dual RNA-seq: major challenges

- Half-life
  - eukaryotic mRNAs half-lives: range of many hours
  - bacterial mRNAs half-life: range of a few minutes (much shorter)

Dual RNA-seq: major challenges

- Depletion of rRNA
- Commercial rRNA removal kits frequently have different efficiencies
  -> may add biases
- Not only to eukaryotic and bacterial rRNAs
- but also decreases the final yield of non-rRNA transcripts
- ∴ ideally rRNA should not be depleted

Dual RNA-seq: major challenges

- sequencing depth
- An estimated minimum of 2,000 million reads from total RNA and 200 million reads from rRNA-depleted samples are required for dual RNA-seq in host and pathogen

Conclusion

- RNA-seq:
- powerful tool to access to the transcriptomes of hosts and pathogens
  - increase the annotation of genomes
  - facilitate analysis under different environmental conditions/ with different genetic backgrounds
- novel way to study host-pathogen interactions
  - detects new virulence/ immune response genes
  - reveals infection mechanism in molecular level

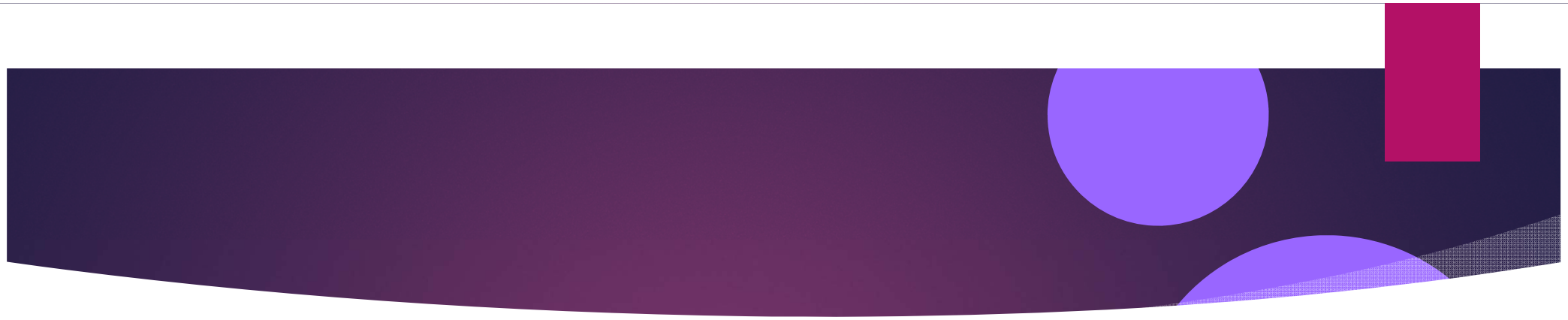
    and in cell type-specific level
  - provides new insight to the clearance of the pathogen

References

- Westermann, A. J., Gorski, S. A., & Vogel, J. (2012). Dual RNA-seq of pathogen and host. *Nature Reviews Microbiology*, *10*(9), 618-630.
- Sultan, M. et al. (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science 321, 956–960.
- Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. & Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 18, 1509–1517.
- Perkins, T. T. et al. (2009). A strand-specific RNA-seq analysis of the transcriptome of the typhoid bacillus Salmonella Typhi. PLoS Genet. 5, e1000569.
- Oliver, H. F. et al. (2009). Deep RNA sequencing of L. monocytogenes reveals overlapping and extensive stationary phase and sigma B-dependent transcriptomes, including multiple highly transcribed noncoding RNAs. BMC Genomics 10, 641.
- Sharma, C. M. et al. (2010). The primary transcriptome of the major human pathogen Helicobacter pylori. Nature 464, 250–255.
- Trapnell, C. et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotech. 28, 511–515.
- Wurtzel, O. et al. (2012). Comparative transcriptomics of pathogenic and non-pathogenic Listeria species. Mol. Syst. Biol. 8, 583.

References

- Bruno, V. M. et al. (2010). Comprehensive annotation of the transcriptome of the human fungal pathogen Candida albicans using RNA-seq. Genome Res. 20, 1451–1458.
- Sorber, K., Dimon, M. T. & DeRisi, J. L. (2011). RNA-Seq analysis of splicing in Plasmodium falciparum uncovers new splice junctions, alternative splicing and splicing of antisense transcripts. Nucleic Acids Res. 39, 3820–3835
- Mandlik, A. et al. (2011). RNA-seq-based monitoring of infection-linked changes in Vibrio cholerae gene expression. Cell Host Microbe 10, 165–174.
- Schulte, L. N., Eulalio, A., Mollenkopf, H. J., Reinhardt, R., & Vogel, J. (2011). Analysis of the host microRNA response to Salmonella uncovers the control of major cytokines by the let-7 family. *The EMBO journal*, *30*(10), 1977-1989.
- Blomström, A. L., Gu, Q., Barry, G., Wilkie, G., Skelton, J. K., Baird, M., ... & Kohl, A. (2015). Transcriptome analysis reveals the host response to Schmallenberg virus in bovine cells and antagonistic effects of the NSs protein. *BMC genomics*, *16*(1), 1.
- Qian, F., Chung, L., Zheng, W., Bruno, V., Alexander, R. P., Wang, Z., ... & Gerstein, M. (2013). Identification of genes critical for resistance to infection by West Nile virus using RNA-Seq analysis. *Viruses*, *5*(7), 1664-1681.
- Liu, Y., Shetty, A. C., Schwartz, J. A., Bradford, L. L., Xu, W., Phan, Q. T., ... & Fraser, C. M. (2015). New signaling pathways govern the host response to C. albicans infection in various niches. *Genome research*, *25*(5), 679-689.
- Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., ... & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome biology*, *17*(1), 1.



End & Thank you

Q&A